FOUNDATIONS OF DESCRIPTIVE AND INFERENTIAL STATISTICS

Lecture notes for the Bachelor degree programmes IB/IMC/IMA/ITM/MEEC/ACM/IEM/IMM

0.1.3 SCIE: Introduction to Scientific Research Methods

0.3.2 RESO: Resources: Financial Resources, Human Resources, Organisation

1.3.2 MARE: Marketing Research

1.4.2 INOP: International Operations

HENK VAN ELST

August 30, 2012

Fakultät I: Betriebswirtschaft und Management
Karlshochschule
International University
Karlsstraße 36–38
76133 Karlsruhe
Germany

E–Mail: hvanelst@karlshochschule.de

Abstract

These lecture notes were written with the aim to provide an accessible though technically solid introduction to the logic of systematical analyses of statistical data to undergraduate and postgraduate students in the Social Sciences and Economics in particular. They may also serve as a general reference for the application of quantitative–empirical research methods. In an attempt to encourage the adoption of an interdisciplinary perspective on quantitative problems arising in practice, the notes cover the four broad topics (i) descriptive statistical processing of raw data, (ii) elementary probability theory, mainly as seen from a frequentist's viewpoint, (iii) the operationalisation of one-dimensional latent variables according to Likert's widely used scaling approach, and (iv) the standard statistical test of hypotheses concerning (a) distributional differences of variables between subgroups of a population, and (b) statistical associations between two variables. The lecture notes are fully hyperlinked, thus providing a direct route to original scientific papers as well as to interesting biographical information. They also list many commands for activating statistical functions and data analysis routines in the software packages SPSS, R and EXCEL/OPEN OFFICE.

These lecture notes were typeset in LaTeX2ε.

Contents

Learning Outcomes for 0.1.3 SCIE

Students who have successfully participated in this module will be able to:

- appropriately apply methods and work techniques of empirical research and adequately implement qualitative and quantitative methods of analysis (e.g. frequency distributions, measures of central tendency, variance and association, correlation between two variables, linear regression).
- understand and describe different approaches to the philosophy of science and epistemology; explain the relationship between the philosophy of science and standards of academic research in the management, economic and social sciences.
- prepare texts, graphs, spreadsheets and presentations using standard software; thereby, be able to communicate in an academically suitable manner as well as convincingly present results.

Learning Outcomes for 0.3.2 RESO

Students who have successfully participated in this module will be able to:

- present the execution of strategic planning within the context of the management process via the selection, procurement, allocation, deployment and organisation of financial and human resources.
- explain the term resources in the context of a “resource-based view”.
- assess, allocate suitably depending on the situation and develop various resources from a general management perspective in the context of varying conditions (“constraints”), strategies and conflict situations (“tensions”).
- apply different methods of researching and making decisions regarding the procurement measures required in a company.
- describe the tasks and instruments of financial management (financial consequences of productivity-based decisions, alternative forms of financing, short and long-term financial and liquidity planning, capital expenditure budgeting including its mathematical principles).
- understand the role of human resource management within the context of general management, explain and critically question the most important structures and processes of HRM and apply selected methods and tools of personnel management.
- present the basic functional, institutional and behaviour-related aspects of the organisation, give a basic outline of research in the field of organisational theory and discuss various theoretical approaches.
- analyse the composition of the organisation and its formal structure, interpret the objectives and conditions of structuring an organisation and assess organisation structures with a view to the situation and cultural context.

Learning Outcomes for 1.3.2 MARE

Students who have successfully participated in this module will be able to:

- gather, record and analyze data in order to identify marketing opportunities and challenges.
- distinguish between market research of the environmental conditions affecting offer and demand and marketing research (of the processes involved in attracting customers and maintaining their loyalty).
- define what is relevant, reliable and valid information for understanding consumer needs.
- appreciate the difference between investigating local consumer needs and those across regional or global markets.
- apply research methods suitable for understanding consumer preferences, attitudes and behaviours in relation to national as well as international contexts; in particular, be able to take into account cultural differences when gathering and interpreting consumer needs in different countries.
- access how changes in the elements of the Marketing Mix affect customer behaviour.

Learning Outcomes for 1.4.2 INOP

Students who have successfully participated in this module will be able to:

- understand how international firms organize their foreign operations.
- comprehend the complexities involved in global sourcing and explain when it is appropriate.
- apply standard concepts, methods and techniques for making decisions on international operations and worldwide logistics.
- apply probability theory and inferential statistics, in order to resolve questions of production planning and control.
- perform sample tests of statistical hypothesis.
- evaluate best practice cases in outsourcing and offshoring.
- analyse current trends in the relocation of productive MNC activities.
- understand the importance of the operations management in order to remain competitive in international markets.

Introductory remarks

Statistical methods of data analysis form the cornerstone of quantitative–empirical research in the **Social Sciences**, **Humanities**, and **Economics**. Historically, the bulk of knowledge available in **Statistics** emerged in the context of the analysis of (large) data sets from observational and experimental measurements in the **Natural Sciences**. The purpose of the present lecture notes is to provide its readers with a solid and thorough, though accessible introduction to the basic concepts of **Descriptive** and **Inferential Statistics**. When discussing methods relating to the latter subject, we will take the perspective of the classical **frequentist approach** to **probability theory**.

The concepts to be introduced and the topics to be covered have been selected in order to make available a fairly self-contained basic statistical tool kit for thorough analysis at the **univariate** and **bivariate** levels of complexity of data gained by means of opinion polls or surveys. In this respect, the present lecture notes are intended to specifically assist the teaching of statistical methods of data analysis in the bachelor degree programmes offered at Karlshochschule International University. In particular, the contents have immediate relevance to solving problems of a quantitative nature in either of (i) the year 1 and year 2 general management modules

- **0.1.3 SCIE: Introduction to Scientific Research Methods**
- **0.3.2 RESO: Resources: Financial Resources, Human Resources, Organisation**,

and (ii) the year 2 special modules of the IB study programme

- **1.3.2 MARE: Marketing Research**
- **1.4.2 INOP: International Operations**.

In the **Social Sciences**, **Humanities**, and **Economics** there are two broad families of empirical research tools available for studying behavioural features of and mutual interactions between human individuals on the one-hand side, and social systems and organisations they form on the other. **Qualitative–empirical methods** focus their view on the individual with the aim to account for her/his/its particular characteristic features, thus probing the "small scale-structure" of a social system, while **quantitative–empirical methods** strive to recognise patterns and regularities that pertain to a large number of individuals and so hope to gain insight on the "large-scale structure" of a social system.

Both approaches are strongly committed to pursuing the principles of the **scientific method**. These entail the systematic observation and measurement of phenomena of interest on the basis of well-defined variables, the structured analysis of data so generated, the attempt to provide compelling

theoretical explanations for effects for which there exists conclusive evidence in the data, the derivation from the data of predictions which can be tested empirically, and the publication of all relevant data and the analytical and interpretational tools developed so that the pivotal reproducibility of a researcher's findings and conclusions is ensured. By complying with these principles, the body of scientific knowledge available in any field of research and in practical applications of scientific insights undergoes a continuing process of updating and expansion.

Having thoroughly worked through these lecture notes, a reader should have obtained a good understanding of the use and efficiency of standard statistical methods for handling quantitative issues as they often arise in a manager's everyday business life. Likewise, a reader should feel well-prepared for a smooth entry into a Master degree programme which puts emphasis on quantitative–empirical methods.

Following a standard pedagogical concept, these lecture notes are split into three main parts: Part I, which comprises Chapters 1 to 5, covers the basic considerations and tools of **Descriptive Statistics**; Part II, which consists of Chapters 6 to 8, introduces the foundations of **Probability Theory**. Finally, the material of Part III, provided in Chapters 9 to 12, first reviews a widespread method for operationalising latent variables, and then introduces a number of standard uni- and bivariate analytical methods of **Inferential Statistics** that prove particularly valuable in practical applications. As such, the contents of Part III are the most important ones for quantitative–empirical research work. Useful mathematical tools have been gathered in an appendix.

Recommended introductory textbooks, which may be used for study in parallel to these lecture notes, are Levin *et al* (2010) [30], Hatzinger and Nagel (2009) [20], Wewel (2008) [60], Toutenburg (2005) [57], or Duller (2007) [11]. These textbooks, as well as many of the monographs listed in the **bibliography**, are available in the library of Karlshochschule International University.

There are *not* included any explicit examples or exercises on the topics discussed. These are reserved to the lectures given throughout term time in any of the modules mentioned.

The present lecture notes are designed to be dynamical in character. On the one-hand side, this means that they will be updated on a regular basis. On the other, they contain interactive features such as fully hyperlinked references, as well as, in the *.pdf version, many active links to biographical references of scientists that have been influential in the historical development of **Probability Theory** and **Statistics**, hosted by the websites The MacTutor History of Mathematics archive (`www-history.mcs.st-and.ac.uk`) and `en.wikipedia.org`.

Lastly, throughout the text references have been provided to respective descriptive and inferential statistical functions and routines that are available on a standard graphic display calculator (GDC), the statistical software packages EXCEL/OPEN OFFICE and SPSS, and, for more technically inclined readers, the widespread statistical software package R. The latter can be obtained as shareware from `cran.r-project.org`, and has been employed for generating the figures included in the text. A useful and easily accessible textbook on the application of R for statistical data analysis is, e.g., Dalgaard (2008) [10]. Further helpful information and assistance is available from the website `www.r-tutor.com`.

Acknowledgments: I am grateful to Kai Holschuh and Eva Kunz for valuable comments on an earlier draft of these lecture notes.

Chapter 1

Statistical variables

A central aim of empirical scientific disciplines is the **observation** of characteristic variable features of a given **system of objects** chosen for study, and the attempt to recognise patterns and regularities which indicate **associations**, or, stronger still, **causal relationships** between them. Based on a combination of **inductive** and **deductive methods of analysis**, the hope is to gain insight of a qualitative and/or quantitative nature into the intricate and often complex interdependencies of such features for the purpose of deriving predictions which can be tested. It is the interplay of experimentation and theoretical modelling, coupled to one another by a number of feedback loops, which generically gives rise to progress in learning and understanding in all scientific activities.

More specifically, the intention is to modify or strengthen the **theoretical foundations** of a scientific discipline by means of observational and/or experimental **falsification** of sets of **hypotheses**. This is generally achieved by employing the quantitative–empirical techniques that have been developed in **Statistics**, in particular in the course of the 20^{th} Century. At the heart of these techniques is the concept of a **statistical variable** X as an entity which represents a single common aspect of the system of objects selected for analysis, the **population** Ω of a **statistical investigation**. In the ideal case, a variable entertains a one-to-one correspondence with an **observable**, and thus is directly amenable to **measurement**. In the **Social Sciences**, **Humanities**, and **Economics**, however, one needs to carefully distinguish between **manifest variables** corresponding to observables on the one-hand side, and **latent variables** representing in general unobservable "social constructs" on the other. It is this latter kind of variables which is commonplace in the fields mentioned. Hence, it becomes necessary to thoroughly address the issue of a reliable, valid and objective **operationalisation** of any given latent variable one has identified as providing essential information on the objects under investigation. A standard approach to dealing with this important matter is reviewed in Ch. 9.

In **Statistics**, it has proven useful to classify variables on the basis of their intrinsic information content into one of three hierachically ordered categories, referred to as **scale levels**. We provide the definition of these scale levels next.

1.1 Scale levels

Def.: Let X be a 1–D **statistical variable** with $k \in \mathbb{N}$ resp. $k \in \mathbb{R}$ possible **values**, **attributes**, or **categorical levels** a_j ($j = 1, \ldots, k$). Statistical variables are classified into one of three hierachically ordered **scale levels of measurement** according to up to three criteria for distinguishing between the possible values or attributes they may take, and the kind of information they contain. One thus defines:

- **Metrically scaled variables** X **(quantitative/numerical)**
  Possible values can be distinguished by

  (i) their *names*, $a_i \neq a_j$,

  (ii) they allow for a *natural ordering*, $a_i < a_j$, and

  (iii) *distances* between them, $a_i - a_j$, are uniquely determined.

  – **Ratio scale**: X has an *absolute zero point* and otherwise only non-negative values; analysis of both differences $a_i - a_j$ and ratios a_i/a_j is meaningful.
  Examples: body height, monthly net income, ...

  – **Interval scale**: X has no *absolute zero point*; only differences $a_i - a_j$ are meaningful.
  Examples: year of birth, temperature in centigrades, ...

- **Ordinally scaled variables** X **(qualitative/categorical)**
  Possible values or attributes can be distinguished by

  (i) their *names*, $a_i \neq a_j$, and

  (ii) they allow for a *natural ordering*, $a_i < a_j$.

  Examples: 5-level Likert item rating scale [Rensis Likert (1903–1981), USA], grading of commodities, ...

- **Nominally scaled variables** X **(qualitative/categorical)**
  Possible values or attributes can be distinguished only by

  (i) their *names*, $a_i \neq a_j$.

  Examples: first name, location of birth, ...

Remark: Note that the applicability of specific methods of **statistical data analysis**, some of which will be discussed in Ch. 11 and 12 below, crucially depends on the **scale level** of the variables involved in the procedures. Metrically scaled variables offer the largest variety of useful methods!

1.2 Raw data sets and data matrices

To set the stage for subsequent considerations, we here introduce some formal representations of entities which assume central roles in statistical data analyses. Let $\mathbf{\Omega}$ denote the **population** of study objects of interest (e.g., human individuals forming a particular social system) relating to some **statistical investigation**. This set $\mathbf{\Omega}$ shall comprise a total of $N \in \mathbb{N}$ **statistical units**, i.e., its size be $|\mathbf{\Omega}| = N$. Suppose one intends to determine the **distributional properties** in $\mathbf{\Omega}$ of $m \in \mathbb{N}$ **statistical variables** $X, Y, \ldots,$ and Z, with **spectra of values** $a_1, a_2, \ldots, a_k$, $b_1, b_2, \ldots, b_l$, $\ldots,$ and $c_1, c_2, \ldots, c_p$, respectively ($k, l, p \in \mathbb{N}$). A **survey** typically obtains from $\mathbf{\Omega}$ a **statistical sample** $\boldsymbol{S_\Omega}$ of size $|\boldsymbol{S_\Omega}| = n$ ($n \in \mathbb{N}$, $n \leq N$), unless one is given the rare opportunity to conduct a proper **census** on $\mathbf{\Omega}$. The data thus generated consists of **observed values** $\{x_i\}_{i=1,\ldots,n}$, $\{y_i\}_{i=1,\ldots,n}$, $\ldots,$ and $\{z_i\}_{i=1,\ldots,n}$. It constitutes the multivariate **raw data set** $\{(x_i, y_i, \ldots, z_i)\}_{i=1,\ldots,n}$ of a statistical investigation and may be conveniently assembled in the form of an $(\boldsymbol{n \times m})$ **data matrix** $\boldsymbol{X}$ given by

sampling unit	**variable** X	**variable** Y	$\ldots$	**variable** Z
1	$x_1 = a_5$	$y_1 = b_9$	$\ldots$	$z_1 = c_3$
2	$x_2 = a_2$	$y_2 = b_{12}$	$\ldots$	$z_2 = c_8$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
n	$x_n = a_8$	$y_n = b_9$	$\ldots$	$z_n = c_{15}$

For recording information obtained from a statistical sample $\boldsymbol{S_\Omega}$, in this matrix scheme every one of the n **sampling units** investigated is allocated a particular row, while every one of the m statistical variables measured is allocated a particular column; in the following, X_{ij} denotes the data entry in the ith row ($i = 1, \ldots, n$) and the jth column ($i = 1, \ldots, m$) of $\boldsymbol{X}$. In general, a $(n \times m)$ data matrix $\boldsymbol{X}$ is the starting point for the application of a statistical software package such as SPSS or R for the purpose of systematic data analysis. Note that in the case of a sample of exclusively metrically scaled data, $\boldsymbol{X} \in \mathbb{R}^{n \times m}$; cf. the lecture notes Ref. [12, Sec. 2.1].

We next turn to describe phenomenologically the **distributional properties** of a single 1–D statistical variable X in a specific statistical sample $\boldsymbol{S_\Omega}$ of size n, drawn in the context of a survey from some population of study objects $\mathbf{\Omega}$ of size N.

Chapter 2

Frequency distributions

The first task at hand in unravelling the intrinsic structure which resides in a given raw data set $\{x_i\}_{i=1,\ldots,n}$ for some statistical variable X corresponds to Cinderella's task of separating the good peas from the bad peas, and collecting them in respective bowls (or bins). This is to say, the first question to be answered requires determination of the **frequencies** with which a value a_j in the spectrum of possible values of X occurred in the statistical sample $\boldsymbol{S_\Omega}$.

2.1 Absolute and relative frequencies

Def.: Let X be a nominally, ordinally or metrically scaled 1–D **statistical variable**, with a spectrum of k different **values** or **attributes** a_j resp. k different **categories** (or bins) K_j ($j = 1, \ldots, k$). If, for X, we have a **raw data set** comprising n **observed values** $\{x_i\}_{i=1,\ldots,n}$, we define by

$$o_j := \begin{cases} o_n(a_j) & = \text{number of } x_i \text{ with } x_i = a_j \\ \\ o_n(K_j) & = \text{number of } x_i \text{ with } x_i \in K_j \end{cases} \tag{2.1}$$

($j = 1, \ldots, k$) the **absolute (observed) frequency** of a_j resp. K_j, and, upon division of the o_j by the sample size n, we define by

$$h_j := \begin{cases} \dfrac{o_n(a_j)}{n} \\ \\ \dfrac{o_n(K_j)}{n} \end{cases} \tag{2.2}$$

($j = 1, \ldots, k$) the **relative frequency** of a_j resp. K_j. Note that for all $j = 1, \ldots, k$, we have $0 \leq o_j \leq n$ with $\sum_{j=1}^{k} o_j = n$, and $0 \leq h_j \leq 1$ with $\sum_{j=1}^{k} h_j = 1$. The k value pairs $(a_j, o_j)_{j=1,\ldots,k}$ resp. $(K_j, o_j)_{j=1,\ldots,k}$ represent the **distribution of absolute frequencies**, the k value pairs $(a_j, h_j)_{j=1,\ldots,k}$ resp. $(K_j, h_j)_{j=1,\ldots,k}$ represent the **distribution of relative frequencies** of the a_j resp. K_j in $\boldsymbol{S_\Omega}$.

EXCEL: `FREQUENCY` (dt.: `HÄUFIGKEIT`)
SPSS: Analyze → Descriptive Statistics → Frequencies ...

Typical graphical representations of **relative frequency distributions**, regularly employed in making results of descriptive statistical data analyses public, are the

- **histogram** for *metrically* scaled data,
- **bar chart** for *ordinally* scaled data,
- **pie chart** for *nominally* scaled data.

It is standard practice in **Statistics** to compile from the relative frequency distribution $(a_j, h_j)_{j=1,\ldots,k}$ resp. $(K_j, h_j)_{j=1,\ldots,k}$ of data for some ordinally or metrically scaled 1–D variable X the associated empirical cumulative distribution function. Hereby it is necessary to distinguish the case of data for a variable with a discrete spectrum of values from the case of data for a variable with a continuous spectrum of values. We will discuss this issue next.

2.2 Empirical cumulative distribution function (discrete data)

Def.: Let X be an ordinally or metrically scaled 1–D statistical variable, the spectrum of values a_j ($j = 1, \ldots, k$) of which vary *discretely*. Suppose given for X a statistical sample $\boldsymbol{S_\Omega}$ of size $|\boldsymbol{S_\Omega}| = n$ comprising observed values $\{x_i\}_{i=1,\ldots,n}$, which we assume ordered in increasing fashion according to $a_1 < a_2 < \ldots < a_k$. The corresponding relative frequency distribution is $(a_j, h_j)_{j=1,\ldots,k}$. For all real numbers $x \in \mathbb{R}$, we then define by

$$F_n(x) := \begin{cases} 0 & \text{for } x < a_1 \\ \sum_{i=1}^{j} h_n(a_i) & \text{for } a_j \leq x < a_{j+1} \qquad (j = 1, \ldots, k-1) \\ 1 & \text{for } x \geq a_k \end{cases} \tag{2.3}$$

the **empirical cumulative distribution function** for X. The value of F_n at $x \in \mathbb{R}$ represents the cumulative relative frequencies of all a_j which are less or equal to x. $F_n(x)$ has the following properties:

- its domain is $D(F_n) = \mathbb{R}$, and its range is $W(F_n) = [0, 1]$; hence, F_n is bounded from above and from below,
- it is continuous from the right and monotonously increasing,
- it is constant on all half-open intervals $[a_j, a_{j+1})$, but exhibits jump discontinuities at all a_{j+1}, of size $h_n(a_{j+1})$, and,
- asymptotically, it behaves as $\lim_{x \to -\infty} F_n(x) = 0$ and $\lim_{x \to +\infty} F_n(x) = 1$.

Computational rules for $F_n(x)$

1. $h(x \leq d) = F_n(d)$
2. $h(x < d) = F_n(d) - h_n(d)$
3. $h(x \geq c) = 1 - F_n(c) + h_n(c)$
4. $h(x > c) = 1 - F_n(c)$
5. $h(c \leq x \leq d) = F_n(d) - F_n(c) + h_n(c)$
6. $h(c < x \leq d) = F_n(d) - F_n(c)$
7. $h(c \leq x < d) = F_n(d) - F_n(c) - h_n(d) + h_n(c)$
8. $h(c < x < d) = F_n(d) - F_n(c) - h_n(d)$,

wherein c denotes an arbitrary **lower bound**, and d denotes an arbitrary **upper bound**, on the argument x of $F_n(x)$.

2.3 Empirical cumulative distribution function (continuous data)

Def.: Let X be a metrically scaled 1–D statistical variable, the spectrum of values of which vary *continuously*, and let observed values $\{x_i\}_{i=1,\ldots,n}$ for X from a statistical sample $\boldsymbol{S_\Omega}$ of size $|\boldsymbol{S_\Omega}| = n$ be binned into k **class intervals** (or bins) K_j ($j = 1, \ldots, k$), of width b_j, with lower boundary u_j and upper boundary o_j. The distribution of relative frequencies of the class intervals be $(K_j, h_j)_{j=1,\ldots,k}$. Then, for all real numbers $x \in \mathbb{R}$,

$$\tilde{F}_n(x) := \begin{cases} 0 & \text{for } x < u_1 \\ \sum_{i=1}^{j-1} h_i + \frac{h_j}{b_j}(x - u_j) & \text{for } x \in K_j \\ 1 & \text{for } x > o_k \end{cases} \tag{2.4}$$

defines the **empirical cumulative distribution function** for X. $\tilde{F}_n(x)$ has the following properties:

- its domain is $D(\tilde{F}_n) = \mathbb{R}$, and its range is $W(\tilde{F}_n) = [0, 1]$; hence, $\tilde{F}_n$ is bounded from above and from below,
- it is continuous and monotonously increasing, and,

- asymptotically, it behaves as $\lim_{x \to -\infty} \tilde{F}_n(x) = 0$ and $\lim_{x \to +\infty} \tilde{F}_n(x) = 1$.

Computational rules for $\tilde{F}_n(x)$

1. $h(x < d) = h(x \leq d) = \tilde{F}_n(d)$
2. $h(x > c) = h(x \geq c) = 1 - \tilde{F}_n(c)$
3. $h(c < x < d) = h(c \leq x < d) = h(c < x \leq d) = h(c \leq x \leq d) = \tilde{F}_n(d) - \tilde{F}_n(c)$,

wherein c denotes an arbitrary **lower bound**, and d denotes an arbitrary **upper bound**, on the argument x of $\tilde{F}_n(x)$.

Our next step is to introduce a set of scale level-dependent standard descriptive measures which characterise specific properties of univariate and bivariate relative frequency distributions of statistical variables X resp. (X, Y).

Chapter 3

Descriptive measures for univariate distributions

There are four families of scale level-dependent standard measures one employs in **Statistics** to describe characteristic properties of univariate relative frequency distributions. We will introduce these in turn. In the following we suppose given from a survey for some 1–D statistical variable X either (i) a raw data set $\{x_i\}_{i=1,\ldots,n}$ of n measured values, or (ii) a relative frequency distribution $(a_j, h_j)_{j=1,\ldots,k}$ resp. $(K_j, h_j)_{j=1,\ldots,k}$.

3.1 Measures of central tendency

Let us begin with the **measures of central tendency** which intend to convey a notion of "middle" or "centre" of a univariate relative frequency distribution.

3.1.1 Mode

The **mode** x_{mod} (nom, ord, metr) of the relative frequency distribution of any 1–D variable X is that value a_j in X's spectrum which occurred with the highest measured relative frequency in a statistical sample $\boldsymbol{S_\Omega}$. Note that the mode does not necessarily take a unique value.

Def.: $h_n(x_{\mathrm{mod}}) \geq h_n(a_j)$ for all $j = 1, \ldots, k$.

EXCEL: `MODE` (dt.: `MODUS.EINF`, `MODALWERT`)
SPSS: Analyze $\rightarrow$ Descriptive Statistics $\rightarrow$ Frequencies ... $\rightarrow$ Statistics ...: Mode

3.1.2 Median

To determine the **median** $\tilde{x}_{0.5}$ (or Q_2) (ord, metr) of the relative frequency distribution of an ordinally or metrically scaled 1–D variable X, it is necessary to first bring the n observed values $\{x_i\}_{i=1,\ldots,n}$ into their natural hierarchical order, i.e., $x_{(1)} \leq x_{(2)} \leq \ldots \leq x_{(n)}$.

Def.: For the sequentially ordered n observed values $\{x_i\}_{i=1,\ldots,n}$, at most 50% have a rank lower or equal to resp. are less or equal to the median value $\tilde{x}_{0.5}$, and at most 50% have a rank higher or equal to resp. are greater or equal to the median value $\tilde{x}_{0.5}$.

(i) Discrete data $\boxed{F_n(\tilde{x}_{0.5}) \geq 0.5}$

$$\tilde{x}_{0.5} = \begin{cases} x_{(\frac{n+1}{2})} & \text{if } n \text{ is odd} \\ \frac{1}{2}[x_{(\frac{n}{2})} + x_{(\frac{n}{2}+1)}] & \text{if } n \text{ is even} \end{cases} . \tag{3.1}$$

(ii) Binned data $\boxed{\tilde{F}_n(\tilde{x}_{0.5}) = 0.5}$

The class interval K_i contains the median value $\tilde{x}_{0.5}$, if $\sum_{j=1}^{i-1} h_j < 0.5$ and $\sum_{j=1}^{i} h_j \geq 0.5$. Then

$$\tilde{x}_{0.5} = u_i + \frac{b_i}{h_i} \left(0.5 - \sum_{j=1}^{i-1} h_j \right) . \tag{3.2}$$

Alternatively, the median of a statistical sample $\boldsymbol{S_\Omega}$ for a continuous variable X with binned data $(K_j, h_j)_{j=1,\ldots,k}$ can be obtained from the associated empirical cumulative distribution function by solving the condition $\tilde{F}_n(\tilde{x}_{0.5}) \stackrel{!}{=} 0.5$ for $\tilde{x}_{0.5}$; cf. Eq. (2.4).[1]

Remark: Note that the value of the median of a relative frequency distribution is fairly insensitive to so-called **outliers** in a statistical sample.

EXCEL: `MEDIAN` (dt.: `MEDIAN`)
SPSS: Analyze → Descriptive Statistics → Frequencies ... → Statistics ...: Median
R: `median(`*variable*`)`

3.1.3 α–Quantile

A generalisation of the median is the concept of the **α–quantile** $\tilde{x}_\alpha$ (ord, metr) of the relative frequency distribution of an ordinally or metrically scaled 1–D variable X. Again, it is necessary to first bring the n observed values $\{x_i\}_{i=1,\ldots,n}$ into their natural hierarchical order, i.e., $x_{(1)} \leq x_{(2)} \leq \ldots \leq x_{(n)}$.

Def.: For the sequentially ordered n observed values $\{x_i\}_{i=1,\ldots,n}$, for given α with $0 < \alpha < 1$ at most $\alpha\times$100% have a rank lower of equal to resp. are less or equal to the α–quantile $\tilde{x}_\alpha$, and at most $(1-\alpha)\times$100% have a rank higher or equal to resp. are greater or equal to the α–quantile $\tilde{x}_\alpha$.

(i) Discrete data $\boxed{F_n(\tilde{x}_\alpha) \geq \alpha}$

$$\tilde{x}_\alpha = \begin{cases} x_{(k)} & \text{if } n\alpha \notin \mathbb{N}, k > n\alpha \\ \frac{1}{2}[x_{(k)} + x_{(k+1)}] & \text{if } k = n\alpha \in \mathbb{N} \end{cases} . \tag{3.3}$$

[1] From a mathematical point of view this amounts to the following problem: consider a straight line which contains the point with coordinates (x_0, y_0) and has non-zero slope $y'(x_0) \neq 0$, i.e., $y = y_0 + y'(x_0)(x - x_0)$. Re-arranging to solve for the variable x then yields $x = x_0 + [y'(x_0)]^{-1}(y - y_0)$.

(ii) Binned data $\boxed{\tilde{F}_n(\tilde{x}_\alpha) = \alpha}$

The class interval K_i contains the α–quantile $\tilde{x}_\alpha$, if $\sum_{j=1}^{i-1} h_j < \alpha$ and $\sum_{j=1}^{i} h_j \geq \alpha$. Then

$$\tilde{x}_\alpha = u_i + \frac{b_i}{h_i}\left(\alpha - \sum_{j=1}^{i-1} h_j\right) . \tag{3.4}$$

Alternatively, an α–quantile of a statistical sample $\boldsymbol{S_\Omega}$ for a continuous variable X with binned data $(K_j, h_j)_{j=1,\ldots,k}$ can be obtained from the associated empirical cumulative distribution function by solving the condition $\tilde{F}_n(\tilde{x}_\alpha) \stackrel{!}{=} \alpha$ for $\tilde{x}_\alpha$; cf. Eq. (2.4).

Remark: The quantiles $\tilde{x}_{0.25}$, $\tilde{x}_{0.5}$, $\tilde{x}_{0.75}$ (also denoted by Q_1, Q_2, Q_3) have special status. They are referred to as the **first quartile** $\rightarrow$ **second quartile (median)** $\rightarrow$ **third quartile** of a relative frequency distribution for an ordinally or a metrically scaled 1–D X and form the core of the **five number summary** of the respective distribution.

EXCEL: `PERCENTILE` (dt.: `QUANTIL.EXKL`, `QUANTIL`)
SPSS: Analyze $\rightarrow$ Descriptive Statistics $\rightarrow$ Frequencies ... $\rightarrow$ Statistics ...: Percentile(s)
R: `quantile(`*variable*`, ` α`)`

3.1.4 Five number summary

The **five number summary** (ord, metr) of the relative frequency distribution of an ordinally or metrically scaled 1–D variable X is a compact compilation of information giving the (i) lowest rank resp. smallest value, (ii) first quartile, (iii) second quartile or median, (iv) third quartile, and (v) highest rank resp. largest value that X takes in a raw data set $\{x_i\}_{i=1,\ldots,n}$ from a statistical sample $\boldsymbol{S_\Omega}$, i.e.,

$$\{x_{(1)}, \tilde{x}_{0.25}, \tilde{x}_{0.5}, \tilde{x}_{0.75}, x_{(n)}\} . \tag{3.5}$$

Alternative notation: $\{Q_0, Q_1, Q_2, Q_3, Q_4\}$.

EXCEL: `MIN`, `QUARTILE`, `MAX` (dt.: `MIN`, `QUARTILE.EXKL`, `QUARTILE`, `MAX`)
SPSS: Analyze $\rightarrow$ Descriptive Statistics $\rightarrow$ Frequencies ... $\rightarrow$ Statistics ...: Quartiles, Minimum, Maximum
R: `quantile(`*variable*`)`

A very convenient graphical method for transparently displaying distributional features of metrically scaled data relating to a five number summary is provided by a **box plot**; see, e.g., Tukey (1977) [59].

All measures of central tendency which now follow are defined exclusively for characterising relative frequency distributions of *metrically scaled 1–D variables* X only.

3.1.5 Arithmetical mean

The best known measure of central tendency is the dimensionful **arithmetical mean** $\bar{x}$ (metr). Given adequate statistical data, it is defined by:

(i) From raw data set:

$$\boxed{\bar{x} := \frac{1}{n}\,(x_1 + \ldots + x_n) =: \frac{1}{n}\sum_{i=1}^{n} x_i\ .} \tag{3.6}$$

(ii) From relative frequency distribution:

$$\bar{x} := a_1 h_n(a_1) + \ldots + a_k h_n(a_k) =: \sum_{j=1}^{k} a_j h_n(a_j)\ . \tag{3.7}$$

Remarks: (i) The value of the arithmetical mean is very sensitive to **outliers**.
(ii) For binned data one selects the midpoint of each class interval K_i to represent the a_j (provided the raw data set is no longer accessible).

EXCEL: `AVERAGE` (dt.: `MITTELWERT`)
SPSS: Analyze $\rightarrow$ Descriptive Statistics $\rightarrow$ Frequencies ... $\rightarrow$ Statistics ...: Mean
R: `mean(`*variable*`)`

3.1.6 Weighted mean

In practice, one also encounters the dimensionful **weighted mean** $\bar{x}_w$ (metr), defined by

$$\boxed{\bar{x}_w := w_1 x_1 + \ldots + w_n x_n =: \sum_{i=1}^{n} w_i x_i\ ;} \tag{3.8}$$

the n **weight factors** $w_1, \ldots, w_n$ need to satisfy the constraints

$$0 \leq w_1, \ldots, w_n \leq 1 \quad \text{and} \quad w_1 + \ldots + w_n = \sum_{i=1}^{n} w_i = 1\ . \tag{3.9}$$

3.2 Measures of variability

The idea behind the **measures of variability** is to convey a notion of the "spread" of data in a given statistical sample $\boldsymbol{S_\Omega}$, technically referred to also as the **dispersion** of the data. As the realisation of this intention requires a well-defined concept of distance, the measures of variability are meaningful for data relating to *metrically scaled 1–D variables* X only. One can distinguish two kinds of such measures: (i) simple 2-data-point measures, and (ii) sophisticated n-data-point measures. We begin with two examples belonging to the first category.

3.2.1 Range

For a raw data set $\{x_i\}_{i=1,\ldots,n}$ of n observed values for X, the dimensionful **range** R (metr) simply expresses the difference between the largest and the smallest value in this set, i.e.,

$$R := x_{(n)} - x_{(1)} \ . \tag{3.10}$$

The basis of this measure is the ordered data set $x_{(1)} \leq x_{(2)} \leq \ldots \leq x_{(n)}$. Alternatively, the range can be denoted by $R = Q_4 - Q_0$.

SPSS: Analyze → Descriptive Statistics → Frequencies ... → Statistics ...: Range

3.2.2 Interquartile range

In the same spirit as the range, the dimensionful **interquartile range** d_Q (metr) is defined as the difference between the third quantile and the first quantile of the relative frequency distribution for some X, i.e.,

$$d_Q := \tilde{x}_{0.75} - \tilde{x}_{0.25} \ . \tag{3.11}$$

Alternatively, this is $d_Q = Q_3 - Q_1$.

3.2.3 Sample variance

The most frequently employed measure of variability in **Statistics** is the dimensionful n-data-point **sample variance** s^2 (metr), and the related sample standard deviation to be discussed below. Given a raw data set $\{x_i\}_{i=1,\ldots,n}$ for X, its spread is essentially quantified in terms of the sum of squared deviations of the n data points from their common mean $\bar{x}$. Due to the algebraic identity

$$(x_1 - \bar{x}) + \ldots + (x_n - \bar{x}) = \sum_{i=1}^{n} (x_i - \bar{x}) = \left(\sum_{i=1}^{n} x_i \right) - n\bar{x} \overset{\text{Eq. (3.6)}}{\equiv} 0 \ ,$$

there are only $n-1$ **degrees of freedom** involved in this measure. The sample variance is defined by:

(i) From raw data set:

$$\boxed{s^2 := \frac{1}{n-1} \left[\, (x_1 - \bar{x})^2 + \ldots + (x_n - \bar{x})^2 \, \right] =: \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2 \ .} \tag{3.12}$$

alternatively, by the **shift theorem**:[2]

$$s^2 = \frac{1}{n-1} \left[\, x_1^2 + \ldots + x_n^2 - n\bar{x}^2 \, \right] = \frac{1}{n-1} \left[\sum_{i=1}^{n} x_i^2 - n\bar{x}^2 \right] \ . \tag{3.13}$$

[2] That is, the algebraic identity $\sum_{i=1}^{n} (x_i - \bar{x})^2 = \sum_{i=1}^{n} \left(x_i^2 - 2x_i\bar{x} + \bar{x}^2 \right) \overset{\text{Eq. (3.6)}}{\equiv} \sum_{i=1}^{n} x_i^2 - \sum_{i=1}^{n} \bar{x}^2 = \sum_{i=1}^{n} x_i^2 - n\bar{x}^2$.

(ii) From relative frequency distribution:

$$\begin{aligned} s^2 &:= \frac{n}{n-1}\left[\,(a_1-\bar{x})^2 h_n(a_1) + \ldots + (a_k-\bar{x})^2 h_n(a_k)\,\right] \\ &=: \frac{n}{n-1}\sum_{j=1}^{k}(a_j-\bar{x})^2\, h_n(a_j)\;. \end{aligned} \tag{3.14}$$

alternatively:

$$\begin{aligned} s^2 &= \frac{n}{n-1}\left[\,a_1^2 h_n(a_1) + \ldots + a_k^2 h_n(a_k) - \bar{x}^2\,\right] \\ &= \frac{n}{n-1}\left[\sum_{j=1}^{k} a_j^2 h_n(a_j) - \bar{x}^2\right]\;. \end{aligned} \tag{3.15}$$

Remarks: (i) We point out that the alternative formulae for a sample variance provided here prove computationally more efficient.
(ii) For binned data, when one selects the midpoint of each class interval K_j to represent the a_j (given the raw data set is no longer accessible), a correction of Eqs. (3.14) and (3.15) by an additional term $(1/12)(n/n-1)\sum_{j=1}^{k} b_j^2 h_j$ becomes necessary, assuming uniformly distributed data within each class intervals K_j of width b_j; cf. Eq. (8.31).

EXCEL: `VAR` (dt.: `VAR.S`, `VARIANZ`)
SPSS: Analyze → Descriptive Statistics → Frequencies ... → Statistics ...: Variance
R: `var(`*variable*`)`

3.2.4 Sample standard deviation

For ease of handling dimensions associated with a metrically scaled 1–D variable X, one defines the dimensionful **sample standard deviation** s (metr) simply as the positive square root of the sample variance, i.e.,

$$s := +\sqrt{s^2}\;, \tag{3.16}$$

such that a measure for the spread of data results which shares the dimension of X and its arithmetical mean $\bar{x}$.

EXCEL: `STDEV` (dt.: `STABW.S`, `STABW`)
SPSS: Analyze → Descriptive Statistics → Frequencies ... → Statistics ...: Std. deviation
R: `sd(`*variable*`)`

3.2.5 Sample coefficient of variation

For ratio scaled 1–D variables X, a dimensionless relative measure of variability is the **sample coefficient of variation** v (metr: ratio), defined by

$$v := \frac{s}{\bar{x}}\;, \quad \text{if} \quad \bar{x} > 0\;. \tag{3.17}$$

3.2.6 Standardisation

Data for metrically scaled 1–D X is amenable to the process of **standardisation**. By this is meant a transformation procedure $X \to Z$, which generates from data for a dimensionful X, with mean $\bar{x}$ and sample standard deviation $s_X > 0$, data for an equivalent dimensionless variable Z according to

$$\boxed{x_i \mapsto z_i := \frac{x_i - \bar{x}}{s_X} \qquad \text{for all } i = 1, \ldots, n\ .} \tag{3.18}$$

For the resultant Z-data, referred to as the ***z*–scores** of the original metrical X-data, this has the convenient consequence that the corresponding arithmetical mean and the sample variance amount to

$$\bar{z} = 0 \quad \text{and} \quad s_Z^2 = 1\ ,$$

respectively.

3.3 Measures of relative distortion

The third family of measures characterising relative frequency distributions of data $\{x_i\}_{i=1,\ldots,n}$ for metrically scaled 1–D variables X, having specific mean $\bar{x}$ and standard deviation s_X, take a **Gaußian normal distribution** with parameter values equal to mean $\bar{x}$ and s_X as a reference case (cf. Sec. 8.5 below). With respect to this *reference distribution*, one defines two kinds of dimensionless **measures of relative distortion** as described in the following.

3.3.1 Skewness

The **skewness** g_1 (metr) is a dimensionless measure to quantify the degree of relative distortion of a given frequency distribution in the *horizontal* direction. For $n > 2$,

$$g_1 := \frac{n}{(n-1)(n-2)} \sum_{i=1}^{n} \left(\frac{x_i - \bar{x}}{s_X} \right)^3\ , \tag{3.19}$$

wherein the observed values $\{x_i\}_{i=1,\ldots,n}$ enter standardised according to Eq. (3.18). Note that $g_1 = 0$ for an exact Gaußian normal distribution.

EXCEL: `SKEW` (dt.: `SCHIEFE`)
SPSS: Analyze → Descriptive Statistics → Frequencies ... → Statistics ...: Skewness
R: `skewness(`*variable*`)`

3.3.2 Excess kurtosis

The **excess kurtosis** g_2 (metr) is a dimensionless measure to quantify the degree of relative distortion of a given frequency distribution in the *vertical* direction. For $n > 3$,

$$g_2 := \frac{n(n+1)}{(n-1)(n-2)(n-3)} \sum_{i=1}^{n} \left(\frac{x_i - \bar{x}}{s_X} \right)^4 - \frac{3(n-1)^2}{(n-2)(n-3)}\ , \tag{3.20}$$

wherein the observed values $\{x_i\}_{i=1,\ldots,n}$ enter standardised according to Eq. (3.18). Note that $g_2 = 0$ for an exact Gaußian normal distribution.

EXCEL: KURT (dt.: KURT)
SPSS: Analyze $\rightarrow$ Descriptive Statistics $\rightarrow$ Frequencies ... $\rightarrow$ Statistics ...: Kurtosis
R: `kurtosis(`*variable*`)`

3.4 Measures of concentration

Finally, for data $\{x_i\}_{i=1,\ldots,n}$ relating to a ratio scaled 1–D variable X, which has a discrete spectrum of values $\{a_j\}_{j=1,\ldots,k}$ or was binned into k different categories $\{K_j\}_{j=1,\ldots,k}$ with respective midpoints a_j, two kinds of **measures of concentration** are commonplace in **Statistics**; one qualitative in nature, the other quantitative.

Begin by defining the **total sum** for the data $\{x_i\}_{i=1,\ldots,n}$ by

$$S := \sum_{i=1}^{n} x_i = \sum_{j=1}^{k} a_j o_n(a_j) \overset{\text{Eq. (3.6)}}{=} n\bar{x} \,, \tag{3.21}$$

where $(a_j, o_n(a_j))_{j=1,\ldots,k}$ is the absolute frequency distribution of the observed values (or categories) of X. Then the **relative proportion** that the value a_j (or the category K_j) takes in S is

$$\frac{a_j o_n(a_j)}{S} = \frac{a_j h_n(a_j)}{\bar{x}} \,. \tag{3.22}$$

3.4.1 Lorenz curve

From the elements introduced in Eqs. (3.21) and (3.22), the US–American economist Max Otto Lorenz (1876–1959) constructed cumulative relative quantities which constitute the coordinates of a so-called **Lorenz curve** representing concentration in the distribution of the ratio scaled 1–D variable X. These coordinates are defined as follows:

- Horizontal axis:

$$k_i := \sum_{j=1}^{i} \frac{o_n(a_j)}{n} = \sum_{j=1}^{i} h_n(a_j) \qquad (i = 1, \ldots, k) \,, \tag{3.23}$$

- Vertical axis:

$$l_i := \sum_{j=1}^{i} \frac{a_j o_n(a_j)}{S} = \sum_{j=1}^{i} \frac{a_j h_n(a_j)}{\bar{x}} \qquad (i = 1, \ldots, k) \,. \tag{3.24}$$

The initial point on a Lorenz curve is generally the coordinate system's origin, $(k_0, l_0) = (0, 0)$, the final point is $(1, 1)$. As a reference to measure concentration in the distribution of X in qualitative terms, one defines a **null concentration curve** as the bisecting line linking $(0, 0)$ to $(1, 1)$. The

Lorenz curve is interpreted as stating that a point on the curve with coordinates (k_i, l_i) represents the fact that $k_i \times 100\%$ of the n statistical units take a share of $l_i \times 100\%$ in the total sum S for the ratio scaled 1–D variable X. Qualitatively, for given data $\{x_i\}_{i=1,\ldots,n}$, the concentration in the distribution of X is the stronger, the larger is the dip of the Lorenz curve relative to the null concentration curve. Note that in addition to the null concentration curve, one can define as a second reference a **maximum concentration curve** such that only the largest value a_k (or category K_k) in the spectrum of values of X takes the full share of 100% in the total sum S for $\{x_i\}_{i=1,\ldots,n}$.

3.4.2 Normalised Gini coefficient

The Italian statistician, demographer and sociologist Corrado Gini (1884–1965) devised a quantitative measure for concentration in the distribution of a ratio scaled 1–D variable X. The dimensionless **normalised Gini coefficient** G_+ (metr: ratio) can be interpreted geometrically as the ratio of areas

$$G_+ := \frac{(\text{area enclosed between Lorenz and null concentration curves})}{(\text{area enclosed between maximum and null concentration curves})} \; . \tag{3.25}$$

Its related computational definition is given by

$$\boxed{G_+ := \frac{n}{n-1}\left[\sum_{i=1}^{k}(k_{i-1}+k_i)\,\frac{a_i o_n(a_i)}{S} - 1\right] \; .} \tag{3.26}$$

Due to normalisation, the range of values is $0 \leq G_+ \leq 1$. Thus, null concentration amounts to $G_+ = 0$, while maximum concentration amounts to $G_+ = 1$.[3]

[3] In September 2012 it was reported (implicitly) in the public press that the coordinates underlying the Lorenz curve describing the distribution of private equity in Germany at the time were $(0.00, 0.00)$, $(0.50, 0.01)$, $(0.90, 0.50)$, and $(1.00, 1.00)$; cf. Ref. [53]. Given that in this case $n \gg 1$, these values amount to a Gini coefficient of $G_+ = 0.64$.

Chapter 4

Descriptive measures of association for bivariate distributions

Now we come to describe and characterise specific features of bivariate frequency distributions, i.e., intrinsic structures of raw data sets $\{(x_i, y_i)\}_{i=1,\ldots,n}$ obtained from statistical samples $\boldsymbol{S_\Omega}$ for 2–D variables (X, Y) from some population of study objects $\boldsymbol{\Omega}$. Let us suppose that the spectrum of values resp. categories of X is $a_1, a_2, \ldots, a_k$, and the spectrum of values resp. categories of Y is $b_1, b_2, \ldots, b_l$, where $k, l \in \mathbb{N}$. Hence, for the bivariate **joint distribution** there exists a total of $k \times l$ possible combinations $\{(a_i, b_j)\}_{i=1,\ldots,k;j=1,\ldots,l}$ of values resp. categories for (X, Y). In the following we will denote associated absolute (observed) frequencies by $o_{ij} := o_n(a_i, b_j)$, and relative frequencies by $h_{ij} := h_n(a_i, b_j)$.

4.1 $(k \times l)$ contingency tables

Consider a raw data set $\{(x_i, y_i)\}_{i=1,\ldots,n}$ for a 2–D variable (X, Y) giving rise to $k \times l$ combinations of values resp. categories $\{(a_i, b_j)\}_{i=1,\ldots,k;j=1,\ldots,l}$. The bivariate joint distribution of observed **absolute frequencies** o_{ij} may be conveniently represented in terms of a $(\boldsymbol{k \times l})$ **contingency table** (or cross tabulation) by

o_{ij}	b_1	b_2	$\ldots$	b_j	$\ldots$	b_l	Σ_j
a_1	o_{11}	o_{12}	$\ldots$	o_{1j}	$\ldots$	o_{1l}	o_{1+}
a_2	o_{21}	o_{22}	$\ldots$	o_{2j}	$\ldots$	o_{2l}	o_{2+}
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
a_i	o_{i1}	o_{i2}	$\ldots$	o_{ij}	$\ldots$	o_{il}	o_{i+}
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
a_k	o_{k1}	o_{k2}	$\ldots$	o_{kj}	$\ldots$	o_{kl}	o_{k+}
Σ_i	o_{+1}	o_{+2}	$\ldots$	o_{+j}	$\ldots$	o_{+l}	n

, (4.1)

where it holds for all $i = 1, \ldots, k$ and $j = 1, \ldots, l$ that

$$0 \leq o_{ij} \leq n \qquad \text{and} \qquad \sum_{i=1}^{k} \sum_{j=1}^{l} o_{ij} = n \; . \tag{4.2}$$

The corresponding **marginal absolute frequencies** of X and of Y are

$$o_{i+} := o_{i1} + o_{i2} + \ldots + o_{ij} + \ldots + o_{il} =: \sum_{j=1}^{l} o_{ij} \tag{4.3}$$

$$o_{+j} := o_{1j} + o_{2j} + \ldots + o_{ij} + \ldots + o_{kj} =: \sum_{i=1}^{k} o_{ij} \; . \tag{4.4}$$

SPSS: Analyze → Descriptive Statistics → Crosstabs ... → Cells ...: Observed

One obtains the related bivariate joint distribution of observed **relative frequencies** h_{ij} following the systematics of Eq. (2.2) to yield

h_{ij}	b_1	b_2	...	b_j	...	b_l	Σ_j
a_1	h_{11}	h_{12}	...	h_{1j}	...	h_{1l}	h_{1+}
a_2	h_{21}	h_{22}	...	h_{2j}	...	h_{2l}	h_{2+}
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
a_i	h_{i1}	h_{i2}	...	h_{ij}	...	h_{il}	h_{i+}
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
a_k	h_{k1}	h_{k2}	...	h_{kj}	...	h_{kl}	h_{k+}
Σ_i	h_{+1}	h_{+2}	...	h_{+j}	...	h_{+l}	1

. (4.5)

Again, it holds for all $i = 1, \ldots, k$ and $j = 1, \ldots, l$ that

$$0 \leq h_{ij} \leq 1 \qquad \text{and} \qquad \sum_{i=1}^{k} \sum_{j=1}^{l} h_{ij} = 1 \; , \tag{4.6}$$

while the **marginal relative frequencies** of X and of Y are

$$h_{i+} := h_{i1} + h_{i2} + \ldots + h_{ij} + \ldots + h_{il} =: \sum_{j=1}^{l} h_{ij} \tag{4.7}$$

$$h_{+j} := h_{1j} + h_{2j} + \ldots + h_{ij} + \ldots + h_{kj} =: \sum_{i=1}^{k} h_{ij} \; . \tag{4.8}$$

On the basis of a $(k \times l)$ contingency table displaying the relative frequencies of the bivariate joint distribution of some 2–D (X, Y), one may define two kinds of related **conditional relative frequency distributions**, namely (i) the conditional distribution of X given Y by

$$h(a_i|b_j) := \frac{h_{ij}}{h_{+j}} \; , \tag{4.9}$$

and (ii) the conditional distribution of Y given X by

$$h(b_j|a_i) := \frac{h_{ij}}{h_{i+}} \; . \tag{4.10}$$

Then, by means of these conditional distributions, a notion of **statistical independence** of variables X and Y is defined to correspond to the simultaneous properties

$$h(a_i|b_j) = h(a_i) = h_{i+} \qquad \text{and} \qquad h(b_j|a_i) = h(b_j) = h_{+j} \; . \tag{4.11}$$

Given these properties hold, it follows from Eqs. (4.9) and (4.10) that

$$h_{ij} = h_{i+} h_{+j} \; . \tag{4.12}$$

4.2 Measures of association for the metrical scale level

Next, specifically consider a raw data set $\{(x_i, y_i)\}_{i=1,\ldots,n}$ from a statistical sample $\boldsymbol{S_\Omega}$ for some metrically scaled 2–D variable (X, Y). The bivariate joint distribution of (X, Y) in this sample can be conveniently represented graphically in terms of a **scatter plot**. Let us now introduce two kinds of measures for the description of specific characteristic features of such distributions.

4.2.1 Sample covariance

The first standard measure characterising bivariate joint distributions of metrically scaled 2–D (X, Y) descriptively is the dimensionful **sample covariance** s_{XY} (metr), defined by

(i) From raw data set:

$$\begin{aligned} s_{XY} &:= \frac{1}{n-1} \left[(x_1 - \bar{x})(y_1 - \bar{y}) + \ldots + (x_n - \bar{x})(y_n - \bar{y}) \right] \\ &=: \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x}) (y_i - \bar{y}) \; , \end{aligned} \tag{4.13}$$

alternatively:

$$\begin{aligned} s_{XY} &= \frac{1}{n-1} \left[x_1 y_1 + \ldots + x_n y_n - n \bar{x} \bar{y} \right] \\ &= \frac{1}{n-1} \left[\sum_{i=1}^{n} x_i y_i - n \bar{x} \bar{y} \right] \; . \end{aligned} \tag{4.14}$$

(ii) From relative frequency distribution:

$$\begin{aligned} s_{XY} &:= \frac{n}{n-1} \left[(a_1 - \bar{x})(b_1 - \bar{y}) h_{11} + \ldots + (a_k - \bar{x})(b_l - \bar{y}) h_{kl} \right] \\ &=: \frac{n}{n-1} \sum_{i=1}^{k} \sum_{j=1}^{l} (a_i - \bar{x}) (b_j - \bar{y}) \, h_{ij} \; , \end{aligned} \tag{4.15}$$

alternatively:

$$\begin{aligned} s_{XY} &= \frac{n}{n-1} \left[a_1 b_1 h_{11} + \ldots + a_k b_l h_{kl} - \bar{x} \bar{y} \right] \\ &= \frac{n}{n-1} \left[\sum_{i=1}^{k} \sum_{j=1}^{l} a_i b_j h_{ij} - \bar{x} \bar{y} \right] \; . \end{aligned} \tag{4.16}$$

Remark: The alternative formulae provided here prove computationally more efficient.

EXCEL: `COVAR` (dt.: `KOVARIANZ.S`)

It is worthwhile to point out that in the research literature it is standard to define for bivariate joint distributions of metrically scaled 2–D (X, Y) a dimensionful symmetric $(\mathbf{2 \times 2})$ **covariance matrix** $\boldsymbol{S}$ according to

$$\boldsymbol{S} := \begin{pmatrix} s_X^2 & s_{XY} \\ s_{XY} & s_Y^2 \end{pmatrix} , \tag{4.17}$$

the components of which are defined by Eqs. (3.12) and (4.13). The determinant of $\boldsymbol{S}$, given by $\det(\boldsymbol{S}) = s_X^2 s_Y^2 - s_{XY}^2$, is positive as long as $s_X^2 s_Y^2 - s_{XY}^2 > 0$ which applies in most practical cases. Then $\boldsymbol{S}$ is regular, and thus a corresponding inverse $\boldsymbol{S}^{-1}$ exists; cf. Ref.[12, Sec. 3.5].

The concept of a regular covariance matrix $\boldsymbol{S}$ and its inverse $\boldsymbol{S}^{-1}$ generalises in a straightforward fashion to the case of multivariate joint distributions of metrically scaled m–D $(X, Y, \ldots, Z)$, where $\boldsymbol{S} \in \mathbb{R}^{m \times m}$ is given by

$$\boldsymbol{S} := \begin{pmatrix} s_X^2 & s_{XY} & \ldots & s_{ZX} \\ s_{XY} & s_Y^2 & \ldots & s_{YZ} \\ \vdots & \vdots & \ddots & \vdots \\ s_{ZX} & s_{YZ} & \ldots & s_Z^2 \end{pmatrix} . \tag{4.18}$$

4.2.2 Bravais and Pearson's sample correlation coefficient

The sample covariance s_{XY} constitutes the basis for the second standard measure characterising bivariate joint distributions of metrically scaled 2–D (X, Y) descriptively, the normalised dimensionless **sample correlation coefficient** r (metr) devised by the French physicist Auguste Bravais (1811–1863) and the English mathematician and statistician Karl Pearson FRS (1857–1936) for the purpose of analysing corresponding raw data $\{(x_i, y_i)\}_{i=1,\ldots,n}$ for the existence of *linear (!!!)* statistical associations. It is defined in terms of the bivariate sample covariance s_{XY} and the univariate sample standard deviations s_X and s_Y by (cf. Bravais (1846) [7] and Pearson (1901) [41])

$$\boxed{r := \frac{s_{XY}}{s_X s_Y} .} \tag{4.19}$$

Due to normalisation, the range of the sample correlation coefficient is $-1 \leq r \leq +1$. The sign of r encodes the **direction** of a correlation. As to interpreting the **strength** of a correlation via the magnitude $|r|$, in practice one typically employs the following qualitative

Rule of thumb:
$0.0 = |r|$: no correlation
$0.0 < |r| < 0.2$: very weak correlation
$0.2 \leq |r| < 0.4$: weak correlation
$0.4 \leq |r| < 0.6$: moderately strong correlation
$0.6 \leq |r| \leq 0.8$: strong correlation

$0.8 \leq |r| < 1.0$: very strong correlation
$1.0 = |r|$: perfect correlation.

EXCEL: `CORREL` (dt.: `KORREL`)
SPSS: Analyze $\rightarrow$ Correlate $\rightarrow$ Bivariate ...: Pearson
R: `cor(`*variable1*`,` *variable2*`, use="complete.obs")`

In addition to Eq. (4.17), it is convenient to define a dimensionless symmetric $(\mathbf{2 \times 2})$ **correlation matrix** $\boldsymbol{R}$ by

$$\boldsymbol{R} := \begin{pmatrix} 1 & r \\ r & 1 \end{pmatrix} , \tag{4.20}$$

which is regular and positive definite as long as $1 - r^2 > 0$. Then its inverse $\boldsymbol{R}^{-1}$ is given by

$$\boldsymbol{R}^{-1} = \frac{1}{1-r^2} \begin{pmatrix} 1 & -r \\ -r & 1 \end{pmatrix} . \tag{4.21}$$

Note that for *non-correlating* metrically scaled variables X and Y, i.e., when $r = 0$, the correlation matrix degenerates to become a unit matrix, $\boldsymbol{R} = \mathbf{1}$.

Again, the concept of a regular and positive definite correlation matrix $\boldsymbol{R}$, with inverse $\boldsymbol{R}^{-1}$, generalises to multivariate joint distributions of metrically scaled m–D $(X, Y, \ldots, Z)$, where $\boldsymbol{R} \in \mathbb{R}^{m \times m}$ is given by[1]

$$\boldsymbol{R} := \begin{pmatrix} 1 & r_{XY} & \ldots & r_{ZX} \\ r_{XY} & 1 & \ldots & r_{YZ} \\ \vdots & \vdots & \ddots & \vdots \\ r_{ZX} & r_{YZ} & \ldots & 1 \end{pmatrix} . \tag{4.22}$$

Note that $\boldsymbol{R}$ is a dimensionless quantity which, hence, is **scale-invariant**; cf. Sec. 8.9.

4.3 Measures of association for the ordinal scale level

At the ordinal scale level, raw data $\{(x_i, y_i)\}_{i=1,\ldots,n}$ for a 2–D variable (X, Y) is not necessarily quantitative in nature. Therefore, in order to be in a position to define a sensible quantitative bivariate measure of statistical association for ordinal variables, one needs to introduce meaningful surrogate data which is numerical. This task is realised by means of defining so-called **ranks**, which are assigned to the original ordinal data according the procedure described in the following.

Begin by establishing amongst the observed values $\{x_i\}_{i=1,\ldots,n}$ resp. $\{y_i\}_{i=1,\ldots,n}$ their natural hierarchical order, i.e.,

$$x_{(1)} \leq x_{(2)} \leq \ldots \leq x_{(n)} \qquad \text{and} \qquad y_{(1)} \leq y_{(2)} \leq \ldots \leq y_{(n)} . \tag{4.23}$$

[1] Given a data matrix $\boldsymbol{X} \in \mathbb{R}^{n \times m}$ for a metrically scaled m–D $(X, Y, \ldots, Z)$, one can show that upon standardisation of the data according to Eq. (3.18), which amounts to a transformation $\boldsymbol{X} \mapsto \boldsymbol{Z} \in \mathbb{R}^{n \times m}$, the correlation matrix can be represented by $\boldsymbol{R} = \dfrac{1}{n-1} \boldsymbol{Z}^T \boldsymbol{Z}$.

Then, every individual x_i resp. y_i is assigned a numerical **rank** which corresponds to its position in the ordered sequences (4.23):

$$x_i \mapsto R(x_i) \; , \quad y_i \mapsto R(y_i) \; , \qquad \text{for all} \qquad i = 1, \ldots, n \; . \tag{4.24}$$

Should there be any "tied ranks" due to equality of some x_i or y_i, one assigns the arithmetical mean of these ranks to all x_i resp. y_i involved in the "tie". Ultimately, by this procedure the entire bivariate raw data undergoes a transformation

$$\{(x_i, y_i)\}_{i=1,\ldots,n} \mapsto \{[R(x_i), R(y_i)]\}_{i=1,\ldots,n} \; , \tag{4.25}$$

yielding n pairs of ranks to numerically represent the original ordinal data.

Given surrogate rank data, the **means of ranks** always amount to

$$\bar{R}(x) \;\; := \;\; \frac{1}{n} \sum_{i=1}^{n} R(x_i) \;\; = \;\; \frac{n+1}{2} \tag{4.26}$$

$$\bar{R}(y) \;\; := \;\; \frac{1}{n} \sum_{i=1}^{n} R(y_i) \;\; = \;\; \frac{n+1}{2} \; . \tag{4.27}$$

The **variances of ranks** are defined in accordance with Eqs. (3.13) and (3.15), i.e.,

$$s^2_{R(x)} \;\; := \;\; \frac{1}{n-1} \left[\sum_{i=1}^{n} R^2(x_i) - n\bar{R}^2(x) \right] \;\; = \;\; \frac{n}{n-1} \left[\sum_{i=1}^{k} R^2(a_i) h_{i+} - \bar{R}^2(x) \right] \tag{4.28}$$

$$s^2_{R(y)} \;\; := \;\; \frac{1}{n-1} \left[\sum_{i=1}^{n} R^2(y_i) - n\bar{R}^2(y) \right] \;\; = \;\; \frac{n}{n-1} \left[\sum_{j=1}^{l} R^2(b_j) h_{+j} - \bar{R}^2(y) \right] \; . \tag{4.29}$$

In addition, to characterise the joint distribution of ranks, a **covariance of ranks** is defined in line with Eqs. (4.14) and (4.16) by

$$\begin{aligned} s_{R(x)R(y)} \;\; &:= \;\; \frac{1}{n-1} \left[\sum_{i=1}^{n} R(x_i) R(y_i) - n\bar{R}(x)\bar{R}(y) \right] \\ &= \;\; \frac{n}{n-1} \left[\sum_{i=1}^{k} \sum_{j=1}^{l} R(a_i) R(b_j) h_{ij} - \bar{R}(x)\bar{R}(y) \right] \; . \end{aligned} \tag{4.30}$$

On this fairly elaborate technical backdrop, the English psychologist and statistician Charles Edward Spearman FRS (1863–1945) defined a dimensionless **rank correlation coefficient** r_S (ord), in analogy to Eq. (4.19), by (cf. Spearman (1904) [51])

$$\boxed{r_S := \frac{s_{R(x)R(y)}}{s_{R(x)} s_{R(y)}} \; .} \tag{4.31}$$

The range of this rank correlation coefficient is $-1 \leq r_S \leq +1$. Again, while the sign of r_S encodes the **direction** of a rank correlation, in interpreting the **strength** of a rank correlation via the magnitude $|r_S|$ one usually employs the qualitative

Rule of thumb:
$0.0 = |r_S|$: no rank correlation
$0.0 < |r_S| < 0.2$: very weak rank correlation
$0.2 \leq |r_S| < 0.4$: weak rank correlation
$0.4 \leq |r_S| < 0.6$: moderately strong rank correlation
$0.6 \leq |r_S| \leq 0.8$: strong rank correlation
$0.8 \leq |r_S| < 1.0$: very strong rank correlation
$1.0 = |r_S|$: perfect rank correlation.

SPSS: Analyze → Correlate → Bivariate ...: Spearman

When *no tied ranks* occur, Eq. (4.31) simplifies to

$$r_S = 1 - \frac{6 \sum_{i=1}^{n} [R(x_i) - R(y_i)]^2}{n(n^2 - 1)} \; . \tag{4.32}$$

4.4 Measures of association for the nominal scale level

Lastly, let us turn to consider the case of quantifying descriptively the degree of statistical association in raw data $\{(x_i, y_i)\}_{i=1,\ldots,n}$ for a nominally scaled 2–D variable (X, Y) with categories $\{(a_i, b_j)\}_{i=1,\ldots,k;j=1,\ldots,l}$. The starting point are the observed absolute resp. relative (cell) frequencies o_{ij} and h_{ij} of the bivariate joint distribution of (X, Y), with marginal frequencies o_{i+} resp. h_{i+} for X and o_{+j} resp. h_{+j} for Y. The **χ^2–statistic** devised by the English mathematical statistician Karl Pearson FRS (1857–1936) rests on the notion of statistical independence of two variables X and Y in that it takes the corresponding formal condition provided by Eq. (4.12) as a reference. A simple algebraic manipulation of this condition obtains

$$h_{ij} = h_{i+} h_{+j} \quad \Rightarrow \quad \frac{o_{ij}}{n} = \frac{o_{i+}}{n} \, \frac{o_{+j}}{n} \quad \overset{\text{multiplication by } n}{\Rightarrow} \quad o_{ij} = \frac{o_{i+} o_{+j}}{n} \; . \tag{4.33}$$

Pearson's descriptive χ^2–statistic (cf. Pearson (1900) [40]) is then defined by

$$\boxed{\chi^2 := \sum_{i=1}^{k} \sum_{j=1}^{l} \frac{\left(o_{ij} - \frac{o_{i+} o_{+j}}{n}\right)^2}{\frac{o_{i+} o_{+j}}{n}} = n \sum_{i=1}^{k} \sum_{j=1}^{l} \frac{(h_{ij} - h_{i+} h_{+j})^2}{h_{i+} h_{+j}} \; ,} \tag{4.34}$$

whose range of values amounts to $0 \leq \chi^2 \leq \max(\chi^2)$, with $\max(\chi^2) := n \, [\min(k, l) - 1]$.

Remark: Provided $\frac{o_{i+} o_{+j}}{n} \geq 5$ for all $i = 1, \ldots, k$ and $j = 1, \ldots, l$, Pearson's χ^2–statistic can be employed for the analysis of statistical associations for 2–D variables (X, Y) of almost all combinations of scale levels.

The problem with Pearson's χ^2–statistic is that, due to its variable spectrum of values, it is not clear how to interpret the **strength** of statistical associations. This shortcoming can, however, be overcome by resorting to the **measure of association** proposed by the Swedish mathematician, actuary, and statistician Harald Cramér (1893–1985), which basically is the result of a special kind

of normalisation of Pearson's measure. Thus, **Cramér's V**, as it has come to be known, is defined by (cf. Cramér (1946) [8])

$$\boxed{V := \sqrt{\frac{\chi^2}{\max(\chi^2)}}\,,} \tag{4.35}$$

with range $0 \leq V \leq 1$. For the interpretation of results, one may now employ the qualitative

Rule of thumb:
$0.0 \leq V < 0.2$: weak association
$0.2 \leq V < 0.6$: moderately strong association
$0.6 \leq V \leq 1.0$: strong association.

SPSS: Analyze $\rightarrow$ Descriptive Statistics $\rightarrow$ Crosstabs $\ldots \rightarrow$ Statistics $\ldots$: Chi-square, Phi and Cramer's V

Chapter 5

Descriptive linear regression analysis

For strongly correlating sample data $\{(x_i, y_i)\}_{i=1,\ldots,n}$ for some metrically scaled 2–D variable (X, Y), i.e., when $0.6 < |r| \leq 1.0$, it is meaningful to construct a mathematical model of the linear quantitative statistical association so diagnosed. The standard method to realise such a model is due to the German mathematician and astronomer Carl Friedrich Gauß (1777–1855) and is known by the name of **descriptive linear regression analysis**; cf. Gauß (1809) [17]. We here restrict our attention to the case of *simple* linear regression which involves data for two variables only.

To be determined is a **best-fit linear model** to given bivariate metrical data $\{(x_i, y_i)\}_{i=1,\ldots,n}$. The linear model in question can be expressed in mathematical terms by

$$\boxed{\hat{y} = a + bx\ ,} \tag{5.1}$$

with unknown regression coefficients ***y*-intercept** a and **slope** b. Gauß' method works as follows.

5.1 Method of least squares

At first, one has to make a choice: assign X the status of an **independent variable**, and Y the status of a **dependent variable** (or vice versa; usually this freedom of choice does exist, unless one is testing a specific functional relationship $y = f(x)$). Then, considering the measured values x_i for X as fixed, to be minimised for the Y-data is the **sum of the squared vertical deviations** of the measured values y_i from the model values $\hat{y}_i = a + bx_i$ associated with an arbitrary straight line through the **cloud of data points** $\{(x_i, y_i)\}_{i=1,\ldots,n}$ in a scatter plot, i.e., the sum

$$S(a, b) := \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} (y_i - a - bx_i)^2\ . \tag{5.2}$$

$S(a, b)$ constitutes a (non-negative) real-valued function of two variables a and b. Hence, determining its (local) **minimum values** entails satisfying (i) the necessary condition of simultaneously *vanishing* first partial derivatives

$$0 \stackrel{!}{=} \frac{\partial S(a, b)}{\partial a}\ , \qquad 0 \stackrel{!}{=} \frac{\partial S(a, b)}{\partial b}\ , \tag{5.3}$$

— this yields a well-determined (2×2) system of linear equations for the unknowns a and b, cf. Ref. [12, Sec. 3.1] —, and (ii) the sufficient condition of a *positive definite* **Hessian matrix** $H(a,b)$ of second partial derivatives

$$H(a,b) := \begin{pmatrix} \dfrac{\partial^2 S(a,b)}{\partial a^2} & \dfrac{\partial^2 S(a,b)}{\partial a \partial b} \\ \dfrac{\partial^2 S(a,b)}{\partial b \partial a} & \dfrac{\partial^2 S(a,b)}{\partial b^2} \end{pmatrix} . \tag{5.4}$$

$H(a,b)$ is referred to as positive definite when all of its eigenvalues are positive; cf. Ref. [12, Sec. 3.6].

5.2 Empirical regression line

It is a fairly straightforward algebraic exercise (see, e.g., Toutenburg (2004) [56, p 141ff]) to show that the values of the unknowns a and b which determine a unique global minimum of $S(a,b)$ amount to

$$\boxed{b = \frac{s_Y}{s_X} r \ , \qquad a = \bar{y} - b\bar{x} \ .} \tag{5.5}$$

These values are referred to as the **least square estimators** for a and b. Note that they are exclusively expressible in terms of familiar univariate and bivariate measures characterising the joint distribution of X and Y.

With the solutions a and b of Eq. (5.5), the resultant **best-fit linear model** is thus

$$\boxed{\hat{y} = \bar{y} + \frac{s_Y}{s_X} r \, (x - \bar{x}) \ .} \tag{5.6}$$

It may be employed for the purpose of generating intrapolating **predictions** of the kind $x \mapsto \hat{y}$ for x-values confined to the interval $[x_{(1)}, x_{(n)}]$.

EXCEL: `SLOPE`, `INTERCEPT` (dt.: `STEIGUNG`, `ACHSENABSCHNITT`)
SPSS: Analyze → Regression → Linear ...

5.3 Coefficient of determination

The quality of any particular simple linear regression model, its **goodness-of-the-fit**, can be quantified by means of the **coefficient of determination** B (metr). This measure is derived starting from the algebraic identity

$$\sum_{i=1}^{n} (y_i - \bar{y})^2 = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2 + \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \ , \tag{5.7}$$

which, upon conveniently re-arranging, leads to defining a quantity

$$\boxed{B := \frac{\sum_{i=1}^{n}(y_i - \bar{y})^2 - \sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} = \frac{\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}\ ,} \tag{5.8}$$

with range $0 \leq B \leq 1$. For a perfect fit $B = 1$, while for no fit $B = 0$. The coefficient of determination provides a descriptive measure for the proportion of variability of Y in a data set $\{(x_i, y_i)\}_{i=1,\ldots,n}$ that can be accounted for as due to the association with X via the simple linear regression model. Note that in simple linear regression it holds that

$$B = r^2\ ; \tag{5.9}$$

see, e.g., Toutenburg (2004) [56, p 150f]).

EXCEL: `RSQ` (dt.: `BESTIMMTHEITSMASS`)
SPSS: Analyze → Regression → Linear ...

This concludes Part I of these lecture notes: the discussion on **descriptive statistical methods** of **data analysis**. To set the stage for the application of inferential statistical methods in Part III, we now turn to review the elementary concepts underlying **probability theory**.

Chapter 6

Elements of probability theory

All examples of **inferential statistical methods** of **data analysis** to be presented in Chs. 11 and 12 have been developed in the context of the so-called **frequentist approach** to **probability theory**. The frequentist approach was pioneered by the French lawyer and amateur mathematician Pierre de Fermat (1601–1665), the French mathematician, physicist, inventor, writer and Catholic philosopher Blaise Pascal (1623–1662), the Swiss mathematician Jacob Bernoulli (1654–1705), and the French mathematician and astronomer Marquis Pierre Simon de Laplace (1749–1827). It is deeply rooted in the assumption that any particular random experiment can be repeated arbitrarily often under the "same conditions" and completely "independently of one another", so that a theoretical basis for defining "*objective probabilities*" of random events is given via the relative frequencies of very long sequences of repetition of the same random experiment.[1] This is a highly idealised viewpoint, however, which shares only a limited amount of similarity with the actual conditions pertaining to an experimentor's reality.

Not everyone in **Statistics** is entirely happy, though, with the philosophy underlying the frequentist approach to introducing the concept of **probability**. A complementary viewpoint is taken by the framework which originated from the work of the English mathematician and Presbyterian minister Thomas Bayes (1702–1761), and later of Laplace, and so is commonly referred to as the **Bayes–Laplace approach**; cf. Bayes (1763) [1] and Laplace (1812) [27]. A striking qualitative difference to the frequentist approach consists in the use of prior "*subjective probabilities*" for random events representing a persons's individual degree-of-belief in their likelihood, which are subsequently updated by analysing relevant empirical sample data. A discussion of the pros and cons of both approaches to **probability theory** can be found in, e.g., Sivia and Skilling (2006) [47, p 8ff] or Gilboa (2009) [18, Sec. 5.3].

In the following we turn to discuss the general principles on which **probability theory** is built.

6.1 Random events

We begin by introducing some basic formal constructions:

[1] A special role in the context of the frequentist approach to probability theory is assumed by Jacob Bernoulli's law of large numbers, as well as the concept of independently and identically distributed random variables; we will discuss these issues in Sec. 8.13 below.

- **Random experiments**: Random experiments are experiments which can be repeated arbitrarily often under identical conditions, with **events** (or outcomes) that cannot be predicted with certainty. Well-known simple examples are found amongst games of chance such as rolling dice or playing roulette.

- **Sample space $\mathbf{\Omega} = \{\omega_1, \omega_2, \ldots\}$**: The sample space associated with a random experiment is constituted by the set of all possible **elementary events** (or elementary outcomes) ω_i ($i = 1, 2, \ldots$), signified by their property of mutual exclusivity. The sample space $\mathbf{\Omega}$ of a random experiment may contain either

  (i) a finite number n of elementary events; then $|\mathbf{\Omega}| = n$, or

  (ii) countably many elementary events in the sense of a one-to-one correspondence with the set of natural numbers $\mathbb{N}$, or

  (iii) uncountably may elements in the sense of a one-to-one correspondence with the set of real numbers $\mathbb{R}$, or an open or closed subset thereof.

- **Random events $A, B, \ldots \subseteq \mathbf{\Omega}$**: Random events are formally defined as any kind of subsets of $\mathbf{\Omega}$ that can be formed from the elementary events $\omega_i \in \mathbf{\Omega}$.

- **Certain event $\mathbf{\Omega}$**: The certain event is synonymous with the sample space. When a random experiment is conducted "something will happen for sure."

- **Impossible event $\emptyset = \{\} = \bar{\mathbf{\Omega}}$**: The impossible event is the natural complement to the certain event. When a random experiment is conducted "it is not possible that nothing will happen at all."

- **Event space $Z(\mathbf{\Omega}) := \{A | A \subseteq \mathbf{\Omega}\}$**: The event space, also referred to as the **power set** of $\mathbf{\Omega}$, is the set of all possible subsets (random events!) that can be formed from elementary events $\omega_i \in \mathbf{\Omega}$. Its size (or cardinality) is given by $|Z(\mathbf{\Omega})| = 2^{|\mathbf{\Omega}|}$. The event space $Z(\mathbf{\Omega})$ constitutes a so-called **$\boldsymbol{\sigma}$–algebra** associated with the sample space $\mathbf{\Omega}$; cf. Rinne (2008) [45, p 177]. When $|\mathbf{\Omega}| = n$, i.e., when $\mathbf{\Omega}$ is finite, then $|Z(\mathbf{\Omega})| = 2^n$.

In the formulation of probability theoretical laws and computational rules, the following set operations and identities prove useful.

Set operations

1. $\bar{A} = \mathbf{\Omega} \backslash A$ — complementation of set (or event) A ("not A")

2. $A \backslash B = A \cap \bar{B}$ — formation of the difference of sets (or events) A and B ("A, but not B")

3. $A \cup B$ — formation of the union of sets (or events) A and B ("A or B")

4. $A \cap B$ — formation of the intersection of sets (or events) A and B ("A and B")

Computational rules and identities

1. $A \cup B = B \cup A$ and $A \cap B = B \cap A$ (**commutativity**)
2. $(A \cup B) \cup C = A \cup (B \cup C)$ and $(A \cap B) \cap C = A \cap (B \cap C)$ (**associativity**)
3. $(A \cup B) \cap C = (A \cap C) \cup (B \cap C)$ and $(A \cap B) \cup C = (A \cup C) \cap (B \cup C)$ (**distributivity**)
4. $\overline{A \cup B} = \bar{A} \cap \bar{B}$ and $\overline{A \cap B} = \bar{A} \cup \bar{B}$ (**de Morgan's laws**)

Before addressing the central issue of **probability theory**, we first provide the following important

Def.: Suppose given that the sample space $\mathbf{\Omega}$ of some random experiment is **compact**. Then one understands by a finite **complete partition** of $\mathbf{\Omega}$ a set of $n \in \mathbb{N}$ random events $\{A_1, \ldots, A_n\}$ such that

(i) $A_i \cap A_j = \emptyset$ for $i \neq j$, i.e., they are pairwise disjoint, and

(ii) $\bigcup_{i=1}^{n} A_i = \mathbf{\Omega}$, i.e., their union is identical to the sample space.

6.2 Kolmogorov's axioms of probability theory

It took a fairly long time until, by 1933, a unanimously accepted basis of **probability theory** was established. In part the delay was due to problems with providing a unique definition of **probability** and how it could be measured and interpreted in practice. The situation was resolved only when the Russian mathematician Andrey Nikolaevich Kolmogorov (1903–1987) proposed to discard the intention of providing a unique definition of **probability** altogether, but restrict the issue instead to merely prescribing in an axiomatic fashion a minimum set of essential properties any **probability measure** needs to have in order to be coherent. We now recapitulate the axioms that Kolmogorov put forward; cf. Kolmogoroff (1933) [24].

For given random experiment, let $\mathbf{\Omega}$ be its sample space and $Z(\mathbf{\Omega})$ the associated event space. Then a mapping

$$P : Z(\mathbf{\Omega}) \to \mathbb{R}_{\geq 0} \tag{6.1}$$

defines a **probability measure** with the following properties:

1. for all **random events** $A \in Z(\mathbf{\Omega})$, (**non-negativity**)

$$P(A) \geq 0 \,, \tag{6.2}$$

2. for the **certain event** $\mathbf{\Omega} \in Z(\mathbf{\Omega})$, (**normalisability**)

$$P(\mathbf{\Omega}) = 1 \,, \tag{6.3}$$

3. for all **pairwise disjoint random events** $A_1, A_2, \ldots \in Z(\mathbf{\Omega})$, i.e., $A_i \cap A_j = \emptyset$ for all $i \neq j$, (**σ–additivity**)

$$P\left(\bigcup_{i=1}^{\infty} A_i\right) = P(A_1 \cup A_2 \cup \ldots) = P(A_1) + P(A_2) + \ldots = \sum_{i=1}^{\infty} P(A_i)\ . \tag{6.4}$$

The first two axioms imply the property

$$0 \leq P(A) \leq 1\ , \quad \text{for all} \quad A \in Z(\mathbf{\Omega})\ . \tag{6.5}$$

A less strict version of the third axiom is given by requiring only **finite additivity** of a probability measure. This means it shall possess the property

$$P(A_1 \cup A_2) = P(A_1) + P(A_2)\ , \quad \text{for any two} \quad A_1, A_2 \in Z(\mathbf{\Omega}) \quad \text{with} \quad A_1 \cap A_2 = \emptyset\ . \tag{6.6}$$

The following consequences for random events $A, B, A_1, A_2, \ldots \in Z(\mathbf{\Omega})$ can be derived from Kolmogorov's three axioms of probability theory; cf., e.g., Toutenburg (2005) [57, p 19ff]:

Consequences

1. $P(\bar{A}) = 1 - P(A)$
2. $P(\emptyset) = P(\bar{\mathbf{\Omega}}) = 0$
3. $P(A_1 \cup A_2) = P(A_1) + P(A_2) - P(A_1 \cap A_2)$, occasionally referred to as **convexity** of a probability measure; cf. Gilboa (2009) [18, p 160].
4. If $A \subseteq B$, then $P(A) \leq P(B)$.
5. $P(B) = \sum_{i=1}^{n} P(B \cap A_i)$, provided the $n \in \mathbb{N}$ random events A_i constitute a finite **complete partition** of the sample space $\mathbf{\Omega}$.
6. $P(A \backslash B) = P(A) - P(A \cap B)$.

6.3 Laplacian random experiments

Games of chance with a *finite* number n of possible mutually exclusive elementary outcomes, such as flipping a single coin once, rolling a single dye once, or selecting a single playing card from a deck of 32, belong to the simplest kinds of random experiments. In this context, there exists a frequentists' notion of a unique "*objective probability*" associated with any single possible random event (outcome) that may occur. Such probabilities can be computed according to a straightforward procedure due to the French mathematician and astronomer Marquis Pierre Simon de Laplace (1749–1827). The procedure rests on the assumption that the device generating the random events is a "fair" one.

Consider a random experiment, the n **elementary events** ω_i ($i = 1, \ldots, n$) of which are "equally likely," meaning they are assigned **equal probability**:

$$P(\omega_i) = \frac{1}{|\mathbf{\Omega}|} = \frac{1}{n}\,, \quad \text{for all} \quad \omega_i \in \mathbf{\Omega}\ (i = 1, \ldots, n)\,. \tag{6.7}$$

Random experiments of this nature are referred to as **Laplacian random experiments**.

Def.: For a Laplacian random experiment, the probability of an arbitrary random event $A \in Z(\mathbf{\Omega})$ can be computed according to the rule

$$\boxed{P(A) := \frac{|A|}{|\mathbf{\Omega}|} = \frac{\text{Number of cases favourable to } A}{\text{Number of all possible cases}}\,.} \tag{6.8}$$

The probability measure P here is called a **Laplacian probability measure**.

The systematic counting of the possible outcomes of random experiments in general is the central theme of **combinatorics**. We now briefly address its main considerations.

6.4 Combinatorics

At the heart of combinatorical considerations is the well-known **urn model**. This supposes given an urn containing $N \in \mathbb{N}$ balls that are either

(a) all different and thus can be uniquely distinguished from one another, or

(b) there are $s \in \mathbb{N}$ ($s \leq N$) subsets of indistinguishable like balls, of sizes $n_1, \ldots, n_s$ resp., such that $n_1 + \ldots + n_s = N$.

6.4.1 Permutations

Permutations relate to the number of distinguishable possibilities of arranging N balls in an ordered sequences. Altogether, for cases (a) resp. (b) one finds that there are a total number of

(a) all balls different	(b) s subsets of like balls
$N!$	$\dfrac{N!}{n_1! n_2! \cdots n_s!}$

different possibilities. The **factorial** of a natural number $N \in \mathbb{N}$ is defined by

$$N! := N \times (N-1) \times (N-2) \times \cdots \times 3 \times 2 \times 1\,. \tag{6.9}$$

6.4.2 Combinations and variations

Combinations and **variations** ask for the total number of distinguishable possibilities of selecting from a collection of N balls a sample of size $n \leq N$, while differentiating between cases when

(a) the order in which balls were selected is either neglected or instead accounted for, and

(b) a ball that was selected once either cannot be selected again or indeed can be selected again as often as possible.

These considerations result in the following cases of

	no repetition	with repetition
combinations (order neglected)	$\left(\begin{array}{c} N \\ n \end{array} \right)$	$\left(\begin{array}{c} N+n-1 \\ n \end{array} \right)$
variations (order accounted for)	$\left(\begin{array}{c} N \\ n \end{array} \right) n!$	N^n

different possibilities. Remember, herein, that the **binomial coefficient** for natural numbers $n, N \in \mathbb{N}$, $n \leq N$ is defined by

$$\left(\begin{array}{c} N \\ n \end{array} \right) := \frac{N!}{n!(N-n)!} \, , \tag{6.10}$$

and satisfies the identity

$$\left(\begin{array}{c} N \\ n \end{array} \right) \equiv \left(\begin{array}{c} N \\ N-n \end{array} \right) \, . \tag{6.11}$$

To conclude this chapter, we turn to discuss the important concept of **conditional probabilities** of random events.

6.5 Conditional probabilities

Consider some random experiment with sample space $\mathbf{\Omega}$, event space $Z(\mathbf{\Omega})$, and a well-defined, unique probability measure P.

Def.: For random events $A, B \in Z(\mathbf{\Omega})$, with $P(B) > 0$,

$$\boxed{P(A|B) := \frac{P(A \cap B)}{P(B)}} \tag{6.12}$$

defines the **conditional probability** of A to happen, given that B happened before. Analogously, one defines a conditional probability $P(B|A)$ with the roles of random events A and B switched, provided $P(A) > 0$.

Def.: Random events $A, B \in Z(\mathbf{\Omega})$ are called **mutually stochastically independent**, if, simultaneously, the conditions

$$\boxed{P(A|B) \stackrel{!}{=} P(A)\ , \quad P(B|A) \stackrel{!}{=} P(B) \quad \stackrel{\text{Eq. 6.12}}{\Leftrightarrow} \quad P(A \cap B) = P(A)P(B)} \tag{6.13}$$

are satisfied, i.e., when for both random events A and B the ***a posteriori*** **probabilities** $P(A|B)$ and $P(B|A)$ coincide with the respective ***a priori*** **probabilities** $P(A)$ and $P(B)$.

For applications, the following two prominent laws of **probability theory** prove essential.

6.5.1 Law of total probability

By the **law of total probability** it holds for any random event $B \in Z(\mathbf{\Omega})$ that

$$\boxed{P(B) = \sum_{i=1}^{m} P(B|A_i)P(A_i)\ ,} \tag{6.14}$$

provided the random events $A_1, \ldots, A_m \in Z(\mathbf{\Omega})$ constitute a finite **complete partition** of $\mathbf{\Omega}$ into $m \in \mathbb{N}$ **pairwise disjoint events**.

6.5.2 Bayes' theorem

This important result is due to the English mathematician and Presbyterian minister Thomas Bayes (1702–1761); cf. Bayes (1763) [1]. It states that for any random event $B \in Z(\mathbf{\Omega})$

$$\boxed{P(A_i|B) = \frac{P(B|A_i)P(A_i)}{\sum_{j=1}^{m} P(B|A_j)P(A_j)}\ ,} \tag{6.15}$$

provided the random events $A_1, \ldots, A_m \in Z(\mathbf{\Omega})$ constitute a finite **complete partition** of $\mathbf{\Omega}$ into $m \in \mathbb{N}$ **pairwise disjoint events**, and $P(B) = \sum_{i=1}^{m} P(B|A_i)P(A_i) > 0$.

The different terms in Eq. (6.15) have been given special names:

- $P(A_i)$: **prior probability** of A_i,
- $P(B|A_i)$: **likelihood** of B, given A_i, and
- $P(A_i|B)$: **posterior probability** of A_i, given B.

On the backdrop of some random event B, **Bayes' theorem** thus foremost relates the posterior probability $P(A_i|B)$ of a particular random event A_i to its prior probability $P(A_i)$. This result forms the basis of the frequently encountered empirical practice of updating one's prior "*subjective probability*" of a specific random event by means of adequate experimental or observational data (and corresponding theoretical considerations); see, e.g., Sivia and Skilling (2006) [47, p 5ff]. In particular, **Bayesian statistics** forms a cornerstone in the mathematical modelling of economic agents' choice behaviour under conditions of uncertainty; cf. the brief review by Svetlova and van Elst (2012) [54], and references therein.

Chapter 7

Discrete and continuous random variables

Applications of **inferential statistical methods** (to be discussed in Chs. 11 and 12 below) rest fundamentally on the concept of a probability-dependent quantity arising in the context of **random experiments** which is referred to as a **random variable**. The present chapter aims to provide a basic introduction to the general properties and characteristic features of these quantities. We begin by stating their definition.

Def.: A real-valued **random variable** is defined by a one-to-one mapping

$$X : \mathbf{\Omega} \to W \subseteq \mathbb{R} \tag{7.1}$$

of the sample space $\mathbf{\Omega}$ of some random experiment into a subset W of the real numbers $\mathbb{R}$.

Depending on the nature of the **spectrum of values** of X, we will distinguish in the following between random variables of the **discrete** and of the **continuous** kind.

7.1 Discrete random variables

Discrete random variables are signified by the existence of a finite or countably infinite

Spectrum of values:

$$X \mapsto x \in \{x_1, \ldots, x_n\} \subset \mathbb{R}\,, \quad \text{with} \quad n \in \mathbb{N}\,. \tag{7.2}$$

All values x_i ($i = 1, \ldots, n$) in this spectrum, referred to as possible **realisations** of X, are assigned individual probabilities p_i by a

Probability function:

$$\boxed{P(X = x_i) = p_i \quad \text{for} \quad i = 1, \ldots, n\,,} \tag{7.3}$$

with properties

(i) $0 \leq p_i \leq 1$, and (**non-negativity**)

(ii) $\displaystyle\sum_{i=1}^{n} p_i = 1.$ (**normalisability**)

Specific distributional features of a discrete random variable X deriving from $P(X = x_i)$ are encoded in the associated theoretical

Cumulative distribution function (`cdf`):

$$\boxed{F_X(x) = \texttt{cdf}(x) := P(X \leq x) = \sum_{i|x_i \leq x} P(X = x_i) \; .} \tag{7.4}$$

The `cdf` exhibits the asymptotic behaviour

$$\lim_{x \to -\infty} F_X(x) = 0 \; , \qquad \lim_{x \to +\infty} F_X(x) = 1 \; . \tag{7.5}$$

Information on the central tendency and the variability of a discrete X resides in its

Expectation value and **variance:**

$$\mathrm{E}(X) \;\; := \;\; \sum_{i=1}^{n} x_i P(X = x_i) \tag{7.6}$$

$$\mathrm{Var}(X) \;\; := \;\; \sum_{i=1}^{n} \left(x_i - \mathrm{E}(X)\right)^2 P(X = x_i) \; . \tag{7.7}$$

By the so-called **shift theorem** it holds that the variance may alternatively be obtained from the computationally more efficient formula

$$\mathrm{Var}(X) = \mathrm{E}\left[(X - \mathrm{E}(X))^2\right] = \mathrm{E}(X^2) - \left[\mathrm{E}(X)\right]^2 \; . \tag{7.8}$$

Specific values of $\mathrm{E}(X)$ and $\mathrm{Var}(X)$ will be denoted throughout by the Greek letters μ and σ^2, respectively.

The evaluation of **event probabilities** for a discrete random variable X follows from the

Computational rules:

$$P(X \leq d) \;\; = \;\; F_X(d) \tag{7.9}$$

$$P(X < d) \;\; = \;\; F_X(d) - P(X = d) \tag{7.10}$$

$$P(X \geq c) \;\; = \;\; 1 - F_X(c) + P(X = c) \tag{7.11}$$

$$P(X > c) \;\; = \;\; 1 - F_X(c) \tag{7.12}$$

$$P(c \leq X \leq d) \;\; = \;\; F_X(d) - F_X(c) + P(X = c) \tag{7.13}$$

$$P(c < X \leq d) \;\; = \;\; F_X(d) - F_X(c) \tag{7.14}$$

$$P(c \leq X < d) \;\; = \;\; F_X(d) - F_X(c) - P(X = d) + P(X = c) \tag{7.15}$$

$$P(c < X < d) \;\; = \;\; F_X(d) - F_X(c) - P(X = d) \; , \tag{7.16}$$

where c and d denote arbitrary lower and upper cut-off values imposed on the spectrum of X.

In applications it is frequently of interest to know the values of a discrete `cdf`'s

α–quantiles:
These are realisations x_α of X specifically determined by the condition that X take values $x \leq x_\alpha$ at least with probability α (for $0 < \alpha < 1$), i.e.,

$$F_X(x_\alpha) = P(X \leq x_\alpha) \stackrel{!}{\geq} \alpha \qquad \text{and} \qquad F_X(x) = P(X \leq x) < \alpha \quad \text{for} \quad x < x_\alpha \; . \tag{7.17}$$

7.2 Continuous random variables

Continuous random variables possess an uncountably infinite

Spectrum of values:

$$X \mapsto x \in \mathbb{D} \subseteq \mathbb{R} \, . \tag{7.18}$$

It is therefore no longer meaningful to assign probabilities to individual **realisations** x of X, but only to infinitesimally small intervals $\mathrm{d}x \in \mathbb{D}$ by means of a

Probability density function (`pdf`):

$$\boxed{f_X(x) = \texttt{pdf}(x) \, .} \tag{7.19}$$

Hence, approximately, $P(X \in \mathrm{d}x) \approx f_X(\xi)\,\mathrm{d}x$, for some $\xi \in \mathrm{d}x$. The `pdf` of an arbitrary continuous X has the defining properties:

(i) $f_X(x) \geq 0$ for all $x \in \mathbb{D}$, (**non-negativity**)

(ii) $\displaystyle\int_{-\infty}^{+\infty} f_X(x)\,\mathrm{d}x = 1$, and (**normalisability**)

(iii) $f_X(x) = F'_X(x)$. (**link to `cdf`**)

The evaluation of **event probabilities** for a continuous X rests on the associated theoretical

Cumulative distribution function (`cdf`):

$$\boxed{F_X(x) = \texttt{cdf}(x) := P(X \leq x) = \int_{-\infty}^{x} f_X(t)\,\mathrm{d}t \, .} \tag{7.20}$$

These are to be obtained according to the

Computational rules:

$$P(X = d) = 0 \tag{7.21}$$

$$P(X \leq d) = F_X(d) \tag{7.22}$$

$$P(X \geq c) = 1 - F_X(c) \tag{7.23}$$

$$P(c \leq X \leq d) = F_X(d) - F_X(c) \, , \tag{7.24}$$

where c and d denote arbitrary lower and upper cut-off values imposed on the spectrum of X. Note that, again, the `cdf` exhibits the asymptotic properties

$$\lim_{x \to -\infty} F_X(x) = 0 \, , \qquad \lim_{x \to +\infty} F_X(x) = 1 \, . \tag{7.25}$$

The central tendency and the variabilty of a continuous X are quantified by its

Expectation value and **variance:**

$$\mathrm{E}(X) := \int_{-\infty}^{+\infty} x f_X(x)\,\mathrm{d}x \tag{7.26}$$

$$\mathrm{Var}(X) := \int_{-\infty}^{+\infty} (x - \mathrm{E}(X))^2 \, f_X(x)\,\mathrm{d}x \, . \tag{7.27}$$

Again, by the **shift theorem** the variance may alternatively be obtained from the computationally more efficient formula $\mathrm{Var}(X) = \mathrm{E}\left[(X - \mathrm{E}(X))^2\right] = \mathrm{E}(X^2) - [\mathrm{E}(X)]^2$. Specific values of $\mathrm{E}(X)$ and $\mathrm{Var}(X)$ will be denoted throughout by μ and σ^2, respectively.

The construction of interval estimates for unknown distribution parameters of given populations, and the statistical testing of hypotheses (to be discussed later in Chs. 11 and 12), require explicit knowledge of the **α–quantiles** associated with the `cdf`s of particular continuous random variables X. In the present case, these are defined as follows.

α–quantiles:
X take values $x \leq x_\alpha$ with probability α (for $0 < \alpha < 1$), i.e.,

$$P(X \leq x_\alpha) = F_X(x_\alpha) \stackrel{!}{=} \alpha \quad \overset{F_X(x) \text{ is strictly monotonously increasing}}{\Leftrightarrow} \quad \boxed{x_\alpha = F_X^{-1}(\alpha)}\,. \tag{7.28}$$

Hence, α–quantiles of the distributions of continuous X are determined by the inverse `cdf` F_X^{-1}. For given α, the spectrum of X is thus naturally partitioned into domains $x \leq x_\alpha$ and $x \geq x_\alpha$.

7.3 Lorenz curve for continuous random variables

For the distribution of a continuous random variable X, the associated **Lorenz curve** expressing qualitatively the degree of concentration involved is defined by

$$\boxed{L(x_\alpha) = \frac{\displaystyle\int_{-\infty}^{x_\alpha} t f_X(t)\,\mathrm{d}t}{\displaystyle\int_{-\infty}^{+\infty} t f_X(t)\,\mathrm{d}t}\,,} \tag{7.29}$$

with x_α denoting a particular α–quantile of the distribution in question.

7.4 Linear transformations of random variables

Linear transformations of real-valued random variables X are determined by the two-parameter relation

$$\boxed{Y = a + bX \quad \text{with} \quad a, b \in \mathbb{R}, b \neq 0\,,} \tag{7.30}$$

where Y denotes the resultant new random variable. Transformations of random variables of this kind have the following effects on the computation of expectation values and variances.

7.4.1 Effect on expectation values

1. $\mathrm{E}(a) = a$
2. $\mathrm{E}(bX) = b\mathrm{E}(X)$
3. $\mathrm{E}(Y) = \mathrm{E}(a + bX) = \mathrm{E}(a) + \mathrm{E}(bX) = a + b\mathrm{E}(X)$.

7.4.2 Effect on variances

1. $\mathrm{Var}(a) = 0$
2. $\mathrm{Var}(bX) = b^2\mathrm{Var}(X)$
3. $\mathrm{Var}(Y) = \mathrm{Var}(a + bX) = \mathrm{Var}(a) + \mathrm{Var}(bX) = b^2\mathrm{Var}(X)$.

7.4.3 Standardisation

Standardisation of an arbitrary random variable X, with $\sqrt{\mathrm{Var}(X)} > 0$, implies the determination of a special linear transformation $X \mapsto Z$ according to Eq. (7.30) such that the expectation value and variance of X are re-scaled to their simplest values possible, i.e., $\mathrm{E}(Z) = 0$ and $\mathrm{Var}(Z) = 1$. Hence, the two (in part non-linear) conditions

$$0 \stackrel{!}{=} \mathrm{E}(Z) = a + b\mathrm{E}(X) \quad \text{and} \quad 1 \stackrel{!}{=} \mathrm{Var}(Z) = b^2\mathrm{Var}(X)\,,$$

for unknowns a and b, need to be satisfied simultaneously. These are solved by, respectively,

$$a = -\frac{\mathrm{E}(X)}{\sqrt{\mathrm{Var}(X)}} \quad \text{and} \quad b = \frac{1}{\sqrt{\mathrm{Var}(X)}}\,, \tag{7.31}$$

and so

$$\boxed{X \to Z = \frac{X - \mathrm{E}(X)}{\sqrt{\mathrm{Var}(X)}}\,, \qquad x \mapsto z = \frac{x - \mu}{\sigma} \in \bar{\mathbb{D}} \subseteq \mathbb{R}\,,} \tag{7.32}$$

irrespective of whether the random variable X is of the discrete kind (cf. Sec. 7.1) or of the continuous kind (cf. Sec. 7.2). It is essential for applications to realise that under the process of standardisation the values of event probabilities for a random variable X remain **invariant** (unchanged), i.e.,

$$P(X \leq x) = P\left(\frac{X - \mathrm{E}(X)}{\sqrt{\mathrm{Var}(X)}} \leq \frac{x - \mu}{\sigma}\right) = P(Z \leq z)\,. \tag{7.33}$$

7.5 Sums of random variables and reproductivity

Def.: For a set of n additive random variables $X_1, \ldots, X_n$, one defines a **total sum** random variable Y_n and a **mean** random variable $\bar{X}_n$ according to

$$\boxed{Y_n := \sum_{i=1}^{n} X_i \qquad \text{and} \qquad \bar{X}_n := \frac{1}{n}\,Y_n\,.} \tag{7.34}$$

By *linearity* of the expectation value operation,[1] it then holds that

$$\mathrm{E}(Y_n) = \mathrm{E}\left(\sum_{i=1}^{n} X_i\right) = \sum_{i=1}^{n} \mathrm{E}(X_i) \qquad \text{and} \qquad \mathrm{E}(\bar{X}_n) = \frac{1}{n}\,\mathrm{E}(Y_n)\,. \tag{7.35}$$

[1] That is: $\mathrm{E}(X_1 + X_2) = \mathrm{E}(X_1) + \mathrm{E}(X_2)$.

If, in addition, the $X_1, \ldots, X_n$ are *mutually stochastically independent* according to Eq. (6.13), it follows from Subsec. 7.4.2 that the variances of Y_n and $\bar{X}_n$ are given by

$$\mathrm{Var}(Y_n) = \mathrm{Var}\left(\sum_{i=1}^{n} X_i\right) = \sum_{i=1}^{n} \mathrm{Var}(X_i) \qquad \text{and} \qquad \mathrm{Var}(\bar{X}_n) = \left(\frac{1}{n}\right)^2 \mathrm{Var}(Y_n)\,. \tag{7.36}$$

Def.: Reproductivity of a probability distribution law (`cdf`) $F(x)$ is given when the sum Y_n of n independent and identically distributed (in short: "i.i.d.") additive random variables $X_1, \ldots, X_n$, which each individually satisfy distribution laws $F_{X_i}(x) = F(x)$, inherits *this same* distribution law $F(x)$ from its underlying n random variables. Examples of reproductive distribution laws, to be discussed in the following Ch. 8, are the binomial, the Gaußian normal, and the χ^2–distributions.

Chapter 8

Standard distributions of discrete and continuous random variables

In this chapter, we review (i) the probability distribution laws which one typically encounters as **theoretical distributions** in the context of the **statistical testing of hypotheses** (cf. Chs. 11 and 12), but we also include (ii) cases of well-established pedagogical merit, and (iii) a few examples of rather specialised distribution laws, which, nevertheless, prove to be of interest in the description and modelling of various theoretical market situations in **Economics**. We split our considerations into two main parts according to whether a random variable X underlying a particular distribution law varies discretely or continuously. For each of the cases selected, we list the **spectrum of values** of X, its **probability function** (for discrete X) or **probability density function** (`pdf`) (for continuous X), its **cumulative distribution function** (`cdf`), its **expectation value** and its **variance**, and, in some continuous cases, also its α**–quantiles**. Additional information, e.g. commands by which a specific function may be activated for computational purposes or plotted on a GDC, by EXCEL, or by R, is included where available.

8.1 Discrete uniform distribution

One of the simplest probability distribution laws for a discrete random variable X is given by the one-parameter **discrete uniform distribution**,

$$X \sim L(n) \; , \tag{8.1}$$

which is characterised by the number n of different values in X's

Spectrum of values:

$$X \mapsto x \in \{x_1, \ldots, x_n\} \subset \mathbb{R} \; , \quad \text{with} \quad n \in \mathbb{N} \; . \tag{8.2}$$

Probability function:

$$\boxed{P(X = x_i) = \frac{1}{n} \quad \text{for} \quad i = 1, \ldots, n \; ;} \tag{8.3}$$

its graph is shown in Fig. 8.1 below for $n = 6$.

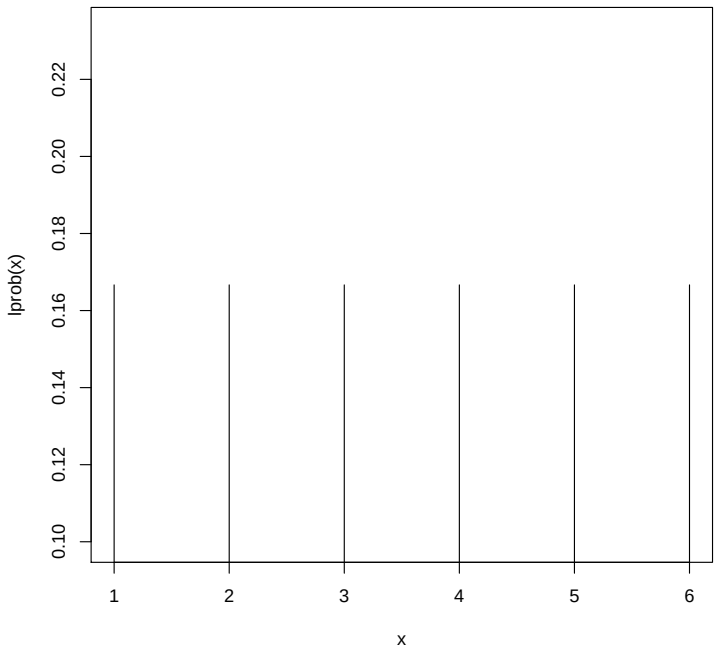

Figure 8.1: Probability function of the discrete uniform distribution according to Eq. (8.3) for the case $L(6)$.

Cumulative distribution function (`cdf`):

$$\boxed{F_X(x) = P(X \leq x) = \sum_{i|x_i \leq x} \frac{1}{n}\,.} \tag{8.4}$$

Expectation value and variance:

$$\mathrm{E}(X) = \sum_{i=1}^{n} x_i \times \frac{1}{n} = \mu \tag{8.5}$$

$$\mathrm{Var}(X) = \left(\sum_{i=1}^{n} x_i^2 \times \frac{1}{n}\right) - \mu^2 = \sigma^2\,. \tag{8.6}$$

The discrete uniform distribution is synonymous with a Laplacian probability measure; cf. Sec. 6.3.

8.2 Binomial distribution

8.2.1 Bernoulli distribution

Another simple probability distribution law, for a discrete random variable X with only two possible values, 0 and 1, is due to the Swiss mathematician Jacob Bernoulli (1654–1705). The **Bernoulli distribution**,

$$X \sim B(1; p) \; , \tag{8.7}$$

depends on a single free parameter, the probability $p \in [0; 1]$ for $X = 1$.

Spectrum of values:

$$X \mapsto x \in \{0, 1\} \; . \tag{8.8}$$

Probability function:

$$\boxed{P(X = x) = \left(\begin{array}{c} 1 \\ x \end{array} \right) p^x (1-p)^{1-x} \; , \quad \text{with} \quad 0 \leq p \leq 1 \; ;} \tag{8.9}$$

its graph is shown in Fig. 8.2 below for $p = \dfrac{1}{3}$.

Cumulative distribution function (`cdf`):

$$\boxed{F_X(x) = P(X \leq x) = \sum_{k=0}^{\lfloor x \rfloor} \left(\begin{array}{c} 1 \\ k \end{array} \right) p^k (1-p)^{1-k} \; .} \tag{8.10}$$

Expectation value and variance:

$$\mathrm{E}(X) \;=\; 0 \times (1-p) + 1 \times p \;=\; p \tag{8.11}$$

$$\mathrm{Var}(X) \;=\; 0^2 \times (1-p) + 1^2 \times p - p^2 \;=\; p(1-p) \; . \tag{8.12}$$

8.2.2 General binomial distribution

A direct generalisation of the Bernoulli distribution is the case of a discrete random variable X which is the *sum* of n mutually stochastically independent, identically Bernoulli-distributed ("i.i.d.") random variables $X_i \sim B(1; p)$ ($i = 1, \ldots, n$), i.e.,

$$X := \sum_{i=1}^{n} X_i = X_1 + \ldots + X_n \; , \tag{8.13}$$

which yields the reproductive two-parameter **binomial distribution**

$$X \sim B(n; p) \; , \tag{8.14}$$

again with $p \in [0; 1]$ the probability for a single $X_i = 1$.

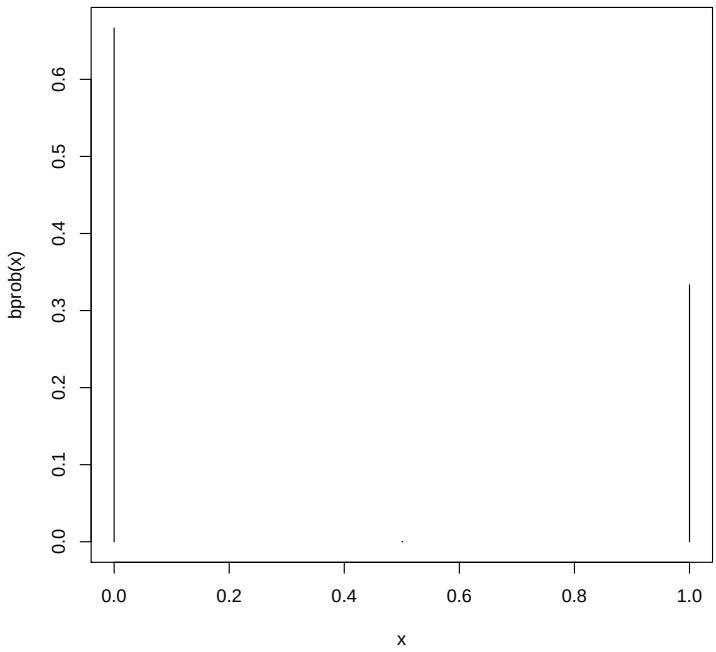

Figure 8.2: Probability function of the Bernoulli distribution according to Eq. (8.9) for the case $B\left(1; \frac{1}{3}\right)$.

Spectrum of values:

$$X \mapsto x \in \{0, \ldots, n\} \ , \quad \text{with} \quad n \in \mathbb{N} \, . \tag{8.15}$$

Probability function:[1]

$$\boxed{P(X = x) = \left(\begin{array}{c} n \\ x \end{array} \right) p^x (1-p)^{n-x} \ , \quad \text{with} \quad 0 \leq p \leq 1 \ ;} \tag{8.16}$$

its graph is shown in Fig. 8.3 below for $n = 10$ and $p = \frac{3}{5}$.

[1] In the context of an urn model with M black balls and $N - M$ white balls, and the random selection of n balls from a total of N, with repetition, this probability function can be derived from Laplace's principle of forming the ratio between the "number of favourable cases" and the "number of all possible cases", cf. Eq. (6.8). Thus, $P(X = x) = \dfrac{\left(\begin{array}{c} n \\ x \end{array} \right) M^x (N-M)^{n-x}}{N^n}$, where x denotes the number of black balls drawn, and one substitutes accordingly from the definition $p := M/N$.

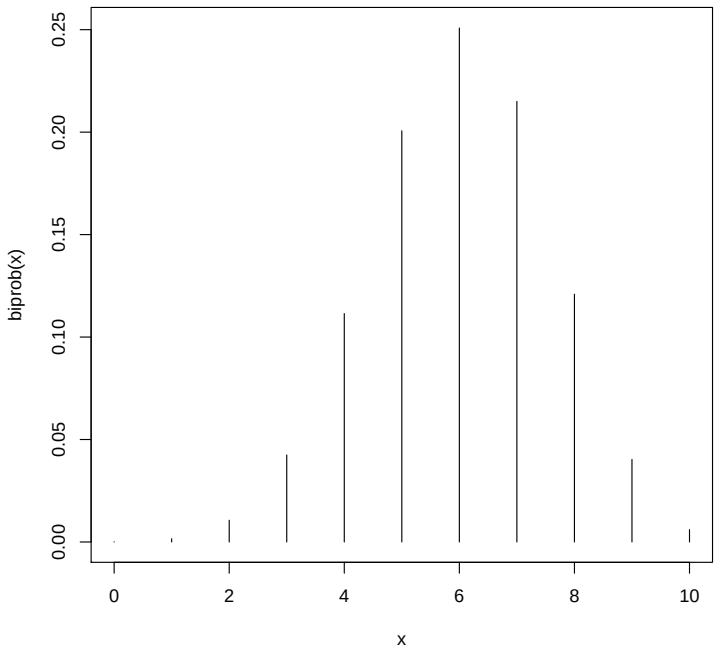

Figure 8.3: Probability function of the binomial distribution according to Eq. (8.16) for the case $B\left(10; \frac{3}{5}\right)$.

Cumulative distribution function (`cdf`):

$$\boxed{F_X(x) = P(X \leq x) = \sum_{k=0}^{\lfloor x \rfloor} \left(\begin{array}{c} n \\ k \end{array} \right) p^k (1-p)^{n-k} \, .} \tag{8.17}$$

Expectation value and variance:

$$\mathrm{E}(X) \;=\; \sum_{i=1}^{n} p \;=\; np \tag{8.18}$$

$$\mathrm{Var}(X) \;=\; \sum_{i=1}^{n} p(1-p) \;=\; np(1-p) \, . \tag{8.19}$$

The results for $\mathrm{E}(X)$ and $\mathrm{Var}(X)$ are based on the rules (7.35) and (7.36), the latter of which applies to a set of mutually stochastically independent random variables.

GDC: `binompdf`(n, p, x), `binomcdf`(n, p, x)
EXCEL: `BINOMDIST` (dt.: `BINOM.VERT`, `BINOMVERT`, `BINOM.INV`)
R: `dbinom`(x, n, p), `pbinom`(x, n, p), `qbinom`(x, n, p)

8.3 Hypergeometric distribution

The **hypergeometric distribution** for a discrete random variable X derives from an urn model with M black balls and $N - M$ white balls, and the random selection of n balls from a total of N ($n \leq N$), without repetition. If X represents the number of black balls amongst the n selected balls, it is subject to the three-parameter probability distribution

$$X \sim H(n, M, N) \; . \tag{8.20}$$

In particular, this model forms the mathematical basis of the internationally popular National Lotteries "6 out of 49", in which case there are $M = 6$ winning numbers amongst the total of $N = 49$ numbers, and $X \in [0; 6]$ denotes the number of correct guesses marked on an individual player's lottery ticket.

Spectrum of values:

$$X \mapsto x \in \{\max(0, n - (N - M)), \ldots, \min(n, M)\} \; . \tag{8.21}$$

Probability function:

$$\boxed{P(X = x) = \frac{\left(\begin{array}{c} M \\ x \end{array} \right) \left(\begin{array}{c} N - M \\ n - x \end{array} \right)}{\left(\begin{array}{c} N \\ n \end{array} \right)} \; .} \tag{8.22}$$

Cumulative distribution function (`cdf`):

$$\boxed{F_X(x) = P(X \leq x) = \sum_{k=\max(0,n-(N-M))}^{\lfloor x \rfloor} \frac{\left(\begin{array}{c} M \\ k \end{array} \right) \left(\begin{array}{c} N - M \\ n - k \end{array} \right)}{\left(\begin{array}{c} N \\ n \end{array} \right)} \; .} \tag{8.23}$$

Expectation value and variance:

$$\mathrm{E}(X) = n \frac{M}{N} \tag{8.24}$$

$$\mathrm{Var}(X) = n \frac{M}{N} \left(1 - \frac{M}{N}\right) \left(\frac{N - n}{N - 1}\right) \; . \tag{8.25}$$

EXCEL: `HYPGEOMDIST` (dt.: `HYPGEOM.VERT`, `HYPGEOMVERT`)

8.4 Continuous uniform distribution

The simplest example of a probability distribution for a continuous random variable X is the **continuous uniform distribution**,

$$X \sim Rec(a; b) \; , \tag{8.26}$$

also referred to as the **rectangular distribution**. Its two free parameters a and b denote the limits of X's

Spectrum of values:

$$X \mapsto x \in [a, b] \subset \mathbb{R} \, . \tag{8.27}$$

Probability density function (`pdf`):[2]

$$\boxed{f_X(x) = \begin{cases} \dfrac{1}{b-a} & \text{for} \quad x \in [a, b] \\ \\ 0 & \text{otherwise} \end{cases} ;} \tag{8.28}$$

its graph is shown in Fig. 8.4 below for three different combinations of the parameters a and b.

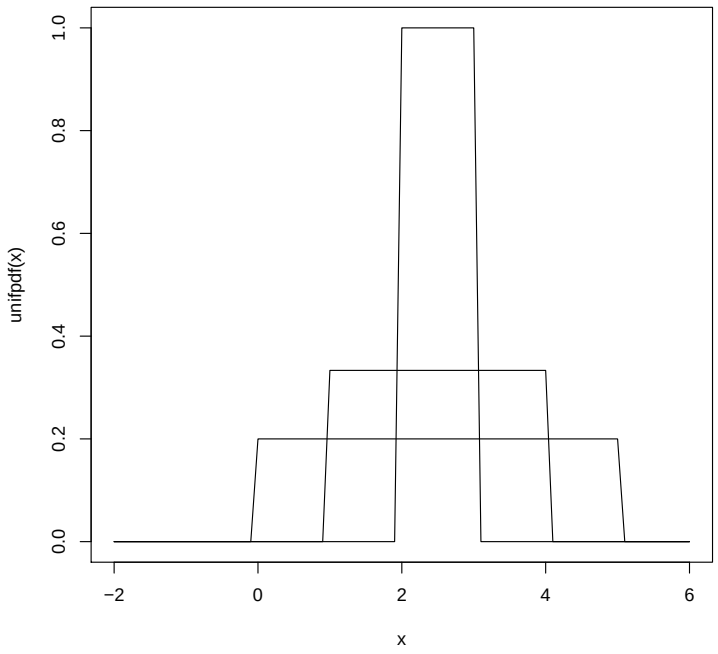

Figure 8.4: `pdf` of the continuous uniform distribution according to Eq. (8.28) for the cases $Rec(0; 5)$, $Rec(1; 4)$ and $Rec(2; 3)$.

[2] It is a nice and instructive little exercise, strongly recommended to the reader, to go through the details of explicitly computing from this simple `pdf` the corresponding `cdf` and the expectation value and variance of $X \sim Rec(a; b)$.

Cumulative distribution function (`cdf`):

$$F_X(x) = P(X \leq x) = \begin{cases} 0 & \text{for} \quad x < a \\ \dfrac{x-a}{b-a} & \text{for} \quad x \in [a,b] \\ 1 & \text{for} \quad x > b \end{cases} . \tag{8.29}$$

Expectation value and variance:

$$\mathrm{E}(X) = \frac{a+b}{2} \tag{8.30}$$

$$\mathrm{Var}(X) = \frac{(b-a)^2}{12} . \tag{8.31}$$

Note that with Eq. (8.29) it is thus a general result that for all continuous uniform distributions

$$\begin{aligned} P(|X - \mathrm{E}(X)| \leq \sqrt{\mathrm{Var}(X)}) &= P\left(\frac{\sqrt{3}(a+b) - (b-a)}{2\sqrt{3}} \leq X \leq \frac{\sqrt{3}(a+b) + (b-a)}{2\sqrt{3}} \right) \\ &= \frac{1}{\sqrt{3}} \approx 0.5773 , \end{aligned} \tag{8.32}$$

i.e., the probability that X falls within one standard deviation ("1–σ") of $\mathrm{E}(X)$ is $1/\sqrt{3}$. α–quantiles of continuous uniform distributions are obtained by straightforward inversion, i.e., for $0 < \alpha < 1$,

$$\alpha \stackrel{!}{=} F_X(x_\alpha) = \frac{x_\alpha - a}{b-a} \qquad \Leftrightarrow \qquad x_\alpha = F_X^{-1}(\alpha) = a + \alpha(b-a) . \tag{8.33}$$

R: `dunif`(x,a,b), `punif`(x,a,b), `qunif`(x,a,b)

Standardisation according to Eq. (7.32) yields

$$X \to Z = \sqrt{3}\,\frac{2X - (a+b)}{b-a} \mapsto z \in \left[-\sqrt{3}, \sqrt{3}\right] , \tag{8.34}$$

with `pdf`

$$f_Z(z) = \begin{cases} \dfrac{1}{2\sqrt{3}} & \text{for} \quad z \in \left[-\sqrt{3}, \sqrt{3}\right] \\ 0 & \text{otherwise} \end{cases} , \tag{8.35}$$

and `cdf`

$$F_Z(z) = P(Z \leq z) = \begin{cases} 0 & \text{for} \quad z < -\sqrt{3} \\ \dfrac{z+\sqrt{3}}{2\sqrt{3}} & \text{for} \quad x \in \left[-\sqrt{3}, \sqrt{3}\right] \\ 1 & \text{for} \quad x > \sqrt{3} \end{cases} . \tag{8.36}$$

8.5 Gaußian normal distribution

The most prominent continuous probability distribution, which is ubiquitous in **Inferential Statistics** (cf. Chs. 11 and 12), is due to Carl Friedrich Gauß (1777–1855): the reproductive two-parameter **normal distribution**

$$X \sim N(\mu; \sigma^2) \; . \tag{8.37}$$

The meaning of the parameters μ and σ^2 will be explained shortly.

Spectrum of values:

$$X \mapsto x \in \mathbb{D} \subseteq \mathbb{R} \; . \tag{8.38}$$

Probability density function (`pdf`):

$$\boxed{f_X(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right] \quad , \quad \text{with} \quad \sigma > 0 \; .} \tag{8.39}$$

This normal–`pdf` defines a reflection-symmetric characteristic bell-shaped curve, the analytical properties of which were first discussed by the French mathematician Abraham de Moivre (1667–1754). The x–position of this curve's (global) maximum is specified by μ, while the x–positions of its two points of inflection are determined by $\mu - \sigma$ resp. $\mu + \sigma$. The effects of different values of the parameters μ and σ on the bell-shaped curve are illustrated in Fig. 8.5 below.

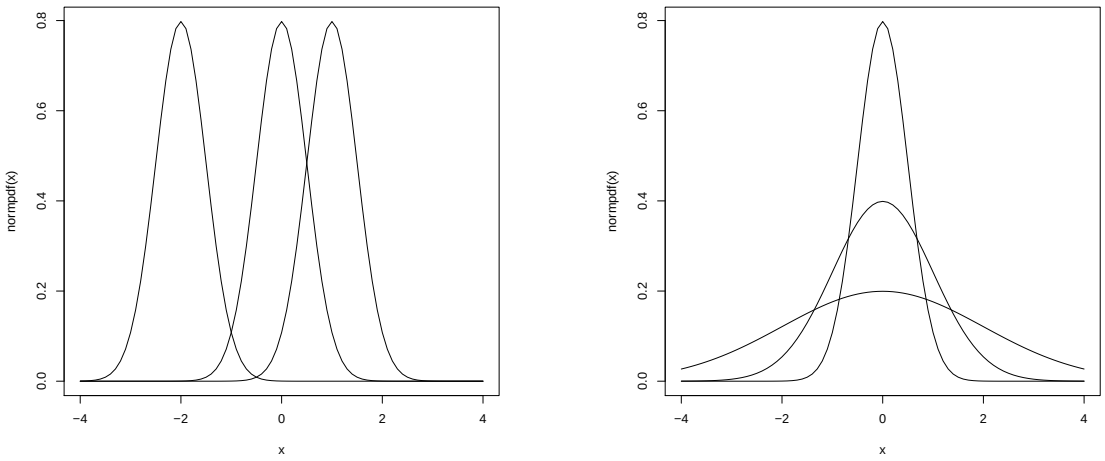

Figure 8.5: `pdf` of the Gaußian normal distribution according to Eq. (8.39). Left panel: cases $N(-2; 1/4)$, $N(0; 1/4)$ and $N(1; 1/4)$. Right panel: cases $N(0; 1/4)$, $N(0; 1)$ and $N(0; 4)$.

Cumulative distribution function (`cdf`):

$$\boxed{F_X(x) = P(X \leq x) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{1}{2}\left(\frac{t-\mu}{\sigma}\right)^2\right] \mathrm{d}t \; .} \tag{8.40}$$

The normal–`cdf` *cannot* be expressed in terms of elementary mathematical functions.

Expectation value and variance:

$$\mathrm{E}(X) = \mu \tag{8.41}$$
$$\mathrm{Var}(X) = \sigma^2 \,. \tag{8.42}$$

GDC: `normalpdf`(x, μ, σ), `normalcdf`$(-\infty, x, \mu, \sigma)$
EXCEL: `NORMDIST` (dt.: `NORM.VERT`, `NORMVERT`)
R: `dnorm`(x, μ, σ), `pnorm`(x, μ, σ)

Upon standardisation of X according to Eq. (7.32), a given normal distribution is transformed into the unique **standard normal distribution**, $Z \sim N(0; 1)$, with

Probability density function (`pdf`):

$$\boxed{\varphi(z) := \frac{1}{\sqrt{2\pi}} \exp\left[-\frac{1}{2}\, z^2\right] \quad \text{for} \quad z \in \mathbb{R}\,;} \tag{8.43}$$

its graph is shown in Fig. 8.6 below.

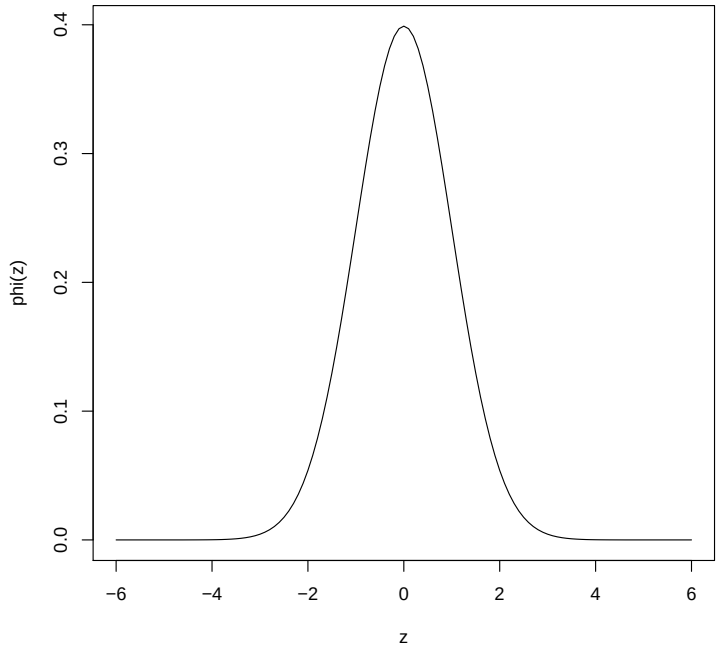

Figure 8.6: `pdf` of the standard normal distribution according to Eq. (8.43).

Cumulative distribution function (`cdf`):

$$\boxed{\Phi(z) := P(Z \le z) = \int_{-\infty}^{z} \frac{1}{\sqrt{2\pi}} \exp\left[-\frac{1}{2}\, t^2\right] \mathrm{d}t \,.} \tag{8.44}$$

The resultant random variable $Z \sim N(0;1)$ satisfies the

Computational rules:

$$\begin{aligned} P(Z \leq b) &= \Phi(b) && (8.45)\\ P(Z \geq a) &= 1 - \Phi(a) && (8.46)\\ P(a \leq Z \leq b) &= \Phi(b) - \Phi(a) && (8.47)\\ \Phi(-z) &= 1 - \Phi(z) && (8.48)\\ P(-z \leq Z \leq z) &= 2\Phi(z) - 1\,. && (8.49) \end{aligned}$$

The probability that a (standard) normally distributed random variable takes values inside an interval of length k times two standard deviations centred on its expectation value is given by the important ***k*σ–rule** according to which

$$P(|X-\mu| \leq k\sigma) \overset{\text{Eq. (7.32)}}{=} P(-k \leq Z \leq +k) \overset{\text{Eq. (8.49)}}{=} 2\Phi(k) - 1 \quad \text{for} \quad k > 0\,. \tag{8.50}$$

α–quantiles associated with $Z \sim N(0;1)$ are obtained from the inverse standard normal–`cdf` according to

$$\alpha \overset{!}{=} P(Z \leq z_\alpha) = \Phi(z_\alpha) \qquad \Leftrightarrow \qquad z_\alpha = \Phi^{-1}(\alpha) \quad \text{for all} \quad 0 < \alpha < 1\,. \tag{8.51}$$

Due to the reflection symmetry of $\varphi(z)$ with respect to the vertical axis at $z = 0$, it holds that

$$z_\alpha = -z_{1-\alpha}\,. \tag{8.52}$$

For this reason, one typically finds z_α listed in textbooks on **Statistics** only for $\alpha \in [1/2, 1)$. Alternatively, a particular z_α may be obtained from a GDC, EXCEL, or R. The backward transformation from a particular z_α of the standard normal distribution to the corresponding x_α of a given normal distribution follows from Eq. (7.32) and amounts to $x_\alpha = \mu + z_\alpha\sigma$.

GDC: `invNorm`(α)
EXCEL: `NORMSINV` (dt.: `NORM.S.INV`, `NORMINV`)
R: `qnorm`(α)

The (standard) normal distribution, as well as the next three examples of continuous probability distributions, are commonly referred to as the **test distributions**, due to the central roles they play in the statistical testing of hypotheses (cf. Chs. 11 and 12).

8.6 χ^2–distribution with n degrees of freedom

The reproductive one-parameter **χ^2–distribution with n degrees of freedom** was devised by the English mathematical statistician Karl Pearson FRS (1857–1936); cf. Pearson (1900) [40]. The underlying continuous random variable

$$\boxed{X \sim \chi^2(n)\,,} \tag{8.53}$$

is perceived of as the sum of squares of n stochastically independent, identically standard normally distributed ("i.i.d.") random variables $Z_i \sim N(0;1)$ ($i = 1, \ldots, n$), i.e.,

$$X := \sum_{i=1}^{n} Z_i^2 = Z_1^2 + \ldots + Z_n^2 \,, \quad \text{with} \quad n \in \mathbb{N} \,. \tag{8.54}$$

Spectrum of values:

$$X \mapsto x \in \mathbb{D} \subseteq \mathbb{R}_{\geq 0} \,. \tag{8.55}$$

The probability density function (`pdf`) of a χ^2–distribution with $df = n$ degrees of freedom is a fairly complicated mathematical expression; see Rinne (2008) [45, p 319] for the explicit representation of the χ^2`pdf`. Plots are shown for four different values of the parameter n in Fig. 8.7. The χ^2`cdf` *cannot* be expressed in terms of elementary mathematical functions.

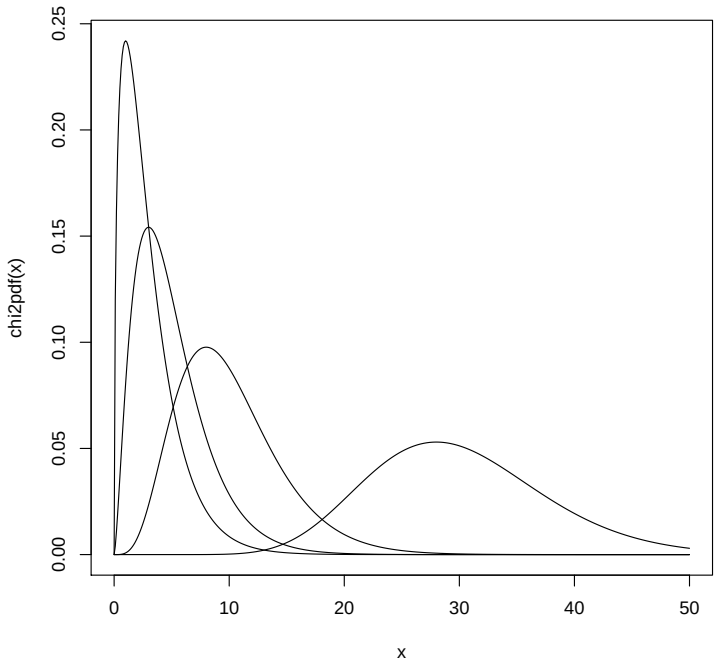

Figure 8.7: `pdf` of the χ^2–distribution for $df = n \in \{3, 5, 10, 30\}$ degrees of freedom. The curves with the highest and lowest peaks correspond to the chases $\chi^2(3)$ and $\chi^2(30)$, respectively.

Expectation value and variance:

$$\mathrm{E}(X) = n \tag{8.56}$$

$$\mathrm{Var}(X) = 2n \,. \tag{8.57}$$

α–quantiles, $\chi^2_{n;\alpha}$, of χ^2–distributions are generally tabulated in textbooks on **Statistics**. Alternatively, they may be obtained from EXCEL or R.

Note that for $n \geq 50$ a $\chi^2(n)$–distribution may be approximated reasonably well by a normal distribution $N(n, 2n)$. This is a reflection of the **central limit theorem**, to be discussed in Sec. 8.13 below.

GDC: χ^2`pdf`(x, n), χ^2`cdf`$(0, x, n)$
EXCEL: `CHIDIST`, `CHIINV` (dt.: `CHIQU.VERT`, `CHIVERT`, `CHIQU.INV`, `CHIINV`)
R: `dchisq`(t, n), `pchisq`(t, n), `qchisq`(t, n)

8.7 t–distribution with n degrees of freedom

The non-reproductive one-parameter ***t*–distribution with n degrees of freedom** was discovered by the English statistician William Sealy Gosset (1876–1937). Somewhat irritating the scientific community, he published his findings under the pseudonym of "Student"; cf. Student (1908) [52]. Consider two stochastically independent random variables, $Z \sim N(0; 1)$ and $X \sim \chi^2(n)$, satisfying the indicated distribution laws. Then the quotient random variable defined by

$$\boxed{T := \frac{Z}{\sqrt{X/n}} \sim t(n)\,, \quad \text{with} \quad n \in \mathbb{N}\,,} \tag{8.58}$$

is t–distributed with $df = n$ degrees of freedom.

Spectrum of values:

$$T \mapsto t \in \mathbb{D} \subseteq \mathbb{R}\,. \tag{8.59}$$

The probability density function (`pdf`) of a t–distribution, which exhibits a reflection symmetry with respect to the vertical axis at $t = 0$, is a fairly complicated mathematical expression; see Rinne (2008) [45, p 326] for the explicit representation of the t`pdf`. Plots are shown for four different values of the parameter n in Fig. 8.8. The t`cdf` *cannot* be expressed in terms of elementary mathematical functions.

Expectation value and variance:

$$\mathrm{E}(X) = 0 \tag{8.60}$$

$$\mathrm{Var}(X) = \frac{n}{n-2} \quad \text{for} \quad n > 2\,. \tag{8.61}$$

α–quantiles, $t_{n;\alpha}$, of t–distributions, for which, due to the reflection symmetry of the t`pdf`, the identity $t_{n;\alpha} = -t_{n;1-\alpha}$ holds, are generally tabulated in textbooks on **Statistics**. Alternatively, they may be obtained from some GDCs, EXCEL, or R.

Note that for $n \geq 50$ a $t(n)$–distribution may be approximated reasonably well by the standard normal distribution, $N(0; 1)$. Again, this is a manifestation of the **central limit theorem**, to be discussed in Sec. 8.13 below.

GDC: `tpdf`(t, n), `tcdf`$(-10, t, n)$, `invT`(α, n)
EXCEL: `TDIST`, `TINV` (dt.: `T.VERT`, `TVERT`, `T.INV`, `TINV`)
R: `dt`(t, n), `pt`(t, n), `qt`(t, n)

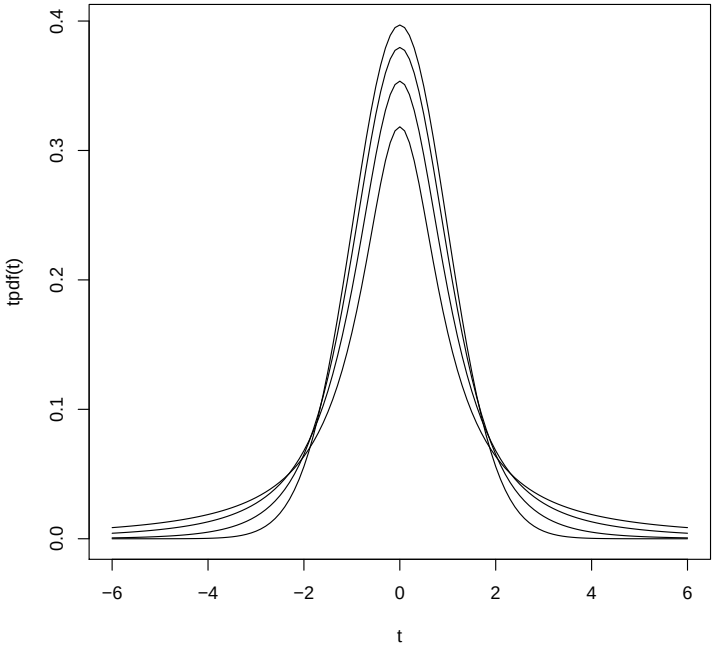

Figure 8.8: `pdf` of the t–distribution for $df = n \in \{1, 2, 5, 50\}$ degrees of freedom. The curves with the lowest and highest peaks correspond to the cases $t(1)$ and $t(50)$, respectively. In the latter, the t`pdf` is essentially equivalent to the standard normal `pdf`. Notice the fatter tails of the t`pdf` for small values of n.

8.8 *F*–distribution with n_1 and n_2 degrees of freedom

The reproductive two-parameter ***F*–distribution with n_1 and n_2 degrees of freedom** was made prominent in **Statistics** by the English statistician, evolutionary biologist, eugenicist and geneticist Sir Ronald Aylmer Fisher FRS (1890–1962), and the US-American mathematician and statistician George Waddel Snedecor (1881–1974); cf. Fisher (1924) [15] and Snedecor (1934) [50]. Consider two sets of stochastically independent, identically standard normally distributed ("i.i.d.") random variables, $X_i \sim N(0;1)$ ($i = 1, \ldots, n_1$), and $Y_j \sim N(0;1)$ ($j = 1, \ldots, n_2$). Define the sums

$$X := \sum_{i=1}^{n_1} X_i^2 \qquad \text{and} \qquad Y := \sum_{j=1}^{n_2} Y_j^2 \,, \tag{8.62}$$

each of which satisfies a χ^2–distribution with n_1 resp. n_2 degrees of freedom. Then the quotient random variable

$$\boxed{F_{n_1,n_2} := \frac{X/n_1}{Y/n_2} \sim F(n_1, n_2) \,, \quad \text{with} \quad n_1, n_2 \in \mathbb{N} \,,} \tag{8.63}$$

is F–distributed with $df_1 = n_1$ and $df_2 = n_2$ degrees of freedom.

Spectrum of values:

$$F_{n_1,n_2} \mapsto f_{n_1,n_2} \in \mathbb{D} \subseteq \mathbb{R}_{\geq 0} \; . \tag{8.64}$$

The probability density function (`pdf`) of an F–distribution is quite a complicated mathematical expression; see Rinne (2008) [45, p 330] for the explicit representation of the F`pdf`. Plots are shown for three different combinations of the parameters n_1 and n_2 in Fig. 8.9. The F`cdf` *cannot* be expressed in terms of elementary mathematical functions.

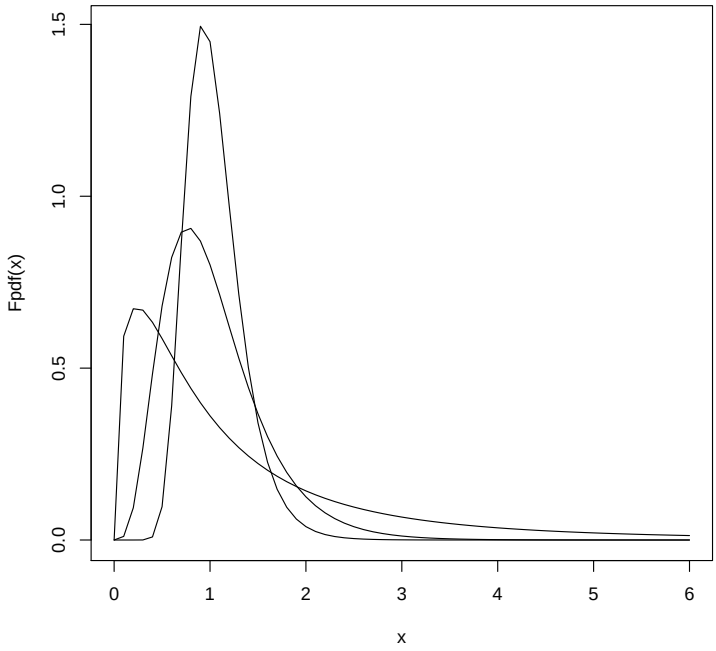

Figure 8.9: `pdf` of the F–distribution for three combinations of degrees of freedom ($df_1 = n_1, df_2 = n_2$). The curves correspond to the cases $F(80, 40)$ (highest peak), $F(10, 50)$, and $F(3, 5)$ (lowest peak), respectively.

Expectation value and variance:

$$\mathrm{E}(X) \;=\; \frac{n_2}{n_2 - 2} \quad \text{for} \quad n_2 > 2 \tag{8.65}$$

$$\mathrm{Var}(X) \;=\; \frac{2n_2^2(n_1 + n_2 - 2)}{n_1(n_2 - 2)^2(n_2 - 4)} \quad \text{for} \quad n_2 > 4 \; . \tag{8.66}$$

α–quantiles, $f_{n_1,n_2;\alpha}$, of F–distributions are tabulated in advanced textbooks on **Statistics**. Alternatively, they may be obtained from EXCEL or R.

GDC: F`pdf`(x, n_1, n_2), F`cdf`$(0, x, n_1, n_2)$
EXCEL: `FDIST`, `FINV` (dt.: `F.VERT`, `FVERT`, `F.INV`, `FINV`)
R: `df`(x, n_1, n_2), `pf`(x, n_1, n_2), `qf`(t, n_1, n_2)

8.9 Pareto distribution

When studying the distribution of wealth and income of people in many different countries at the end of the 19^{th} Century, the Italian engineer, sociologist, economist, political scientist and philosopher Vilfredo Federico Damaso Pareto (1848–1923) discovered certain types of quantitative regularities which he could model mathematically in terms of a simple power-law function involving only two free parameters (cf. Pareto (1896) [39]). The random variable X underlying such a **Pareto distribution**,

$$X \sim Par(\gamma, x_{\min}) \,, \tag{8.67}$$

has a

Spectrum of values:

$$X \mapsto x \in \{x | x \geq x_{\min}\} \subset \mathbb{R}_{>0} \,, \tag{8.68}$$

and a

Probability density function (`pdf`):

$$\boxed{f_X(x) = \begin{cases} 0 & \text{for} \quad x < x_{\min} \\ \\ \dfrac{\gamma}{x_{\min}} \left(\dfrac{x_{\min}}{x}\right)^{\gamma+1} , \quad \gamma \in \mathbb{R}_{>0} & \text{for} \quad x \geq x_{\min} \end{cases} ;} \tag{8.69}$$

its graph is shown in Fig. 8.10 below for three different values of the exponent γ.

Cumulative distribution function (`cdf`):

$$\boxed{F_X(x) = P(X \leq x) = \begin{cases} 0 & \text{for} \quad x < x_{\min} \\ \\ 1 - \left(\dfrac{x_{\min}}{x}\right)^{\gamma} & \text{for} \quad x \geq x_{\min} \end{cases} .} \tag{8.70}$$

Expectation value and variance:

$$\mathrm{E}(X) = \frac{\gamma}{\gamma - 1} x_{\min} \qquad \text{for} \quad \gamma > 1 \tag{8.71}$$

$$\mathrm{Var}(X) = \frac{\gamma}{(\gamma-1)^2(\gamma-2)} x_{\min}^2 \qquad \text{for} \quad \gamma > 2 \,. \tag{8.72}$$

It is important to realise that for $\gamma \leq 1$ neither $\mathrm{E}(X)$ nor $\mathrm{Var}(X)$ are well-defined; for $\gamma \leq 2$ it is only $\mathrm{Var}(X)$ which is not well-defined.

α–quantiles:

$$\alpha \stackrel{!}{=} F_X(x_\alpha) = 1 - \left(\frac{x_{\min}}{x_\alpha}\right)^{\gamma} \;\Leftrightarrow\; x_\alpha = F_X^{-1}(\alpha) = \sqrt[\gamma]{\frac{1}{1-\alpha}}\, x_{\min} \quad \text{for all} \quad 0 < \alpha < 1 \,. \tag{8.73}$$

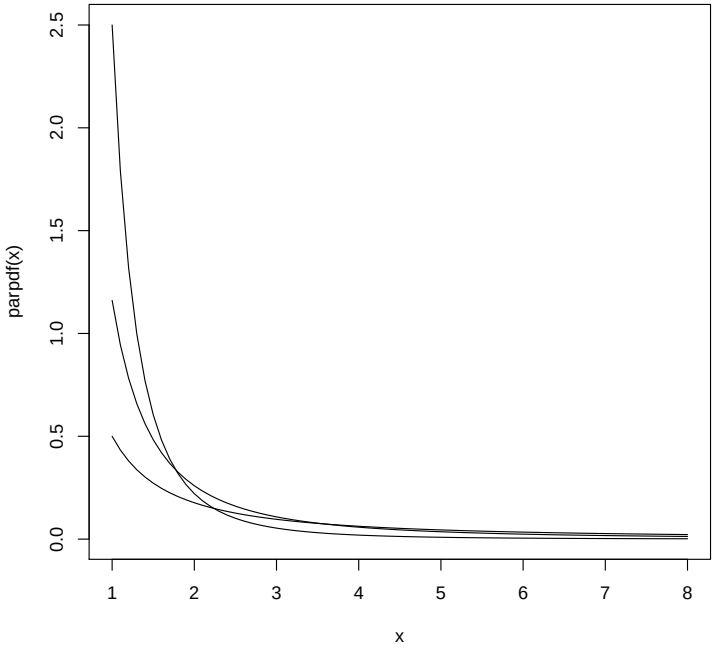

Figure 8.10: `pdf` of the Pareto distribution according to Eq. (8.69) for $x_{\min} = 1$ and $\gamma \in \left\{\frac{1}{2}, \frac{\ln(5)}{\ln(4)}, \frac{5}{2}\right\}$. The curve with the largest value at $x = 1$ corresponds to $Par\left(\frac{5}{2}, 1\right)$.

Note that it follows from Eq. (8.70) that the probability of a Pareto-distributed continuous random variable X to exceed a certain threshold value x is given by the simple power-law rule

$$P(X > x) = 1 - P(X \leq x) = \left(\frac{x_{\min}}{x}\right)^{\gamma} . \tag{8.74}$$

Hence, the ratio of probabilities

$$\frac{P(X > ax)}{P(X > x)} = \frac{\left(\frac{x_{\min}}{ax}\right)^{\gamma}}{\left(\frac{x_{\min}}{x}\right)^{\gamma}} = \left(\frac{1}{a}\right)^{\gamma} , \tag{8.75}$$

with $a \in \mathbb{R}_{>0}$, is **scale-invariant**, meaning independent of a particular scale x at which one observes X (cf. Taleb (2007) [55, p 256ff and p 326ff]). This behaviour is a direct consequence of a special mathematical property of Pareto distributions which is technically referred to as **self-similarity**. It can be defined from the fact that a Pareto-`pdf` (8.69) has *constant* elasticity, i.e. (cf. Ref. [12, Sec. 7.6])

$$\varepsilon_{f_X}(x) = -(\gamma + 1) \quad \text{for} \quad x \geq x_{\min} . \tag{8.76}$$

Further interesting examples of distributions of quantities encountered in various applied fields of science which also feature the scale-invariance of scaling laws are described in Wiesenfeld

(2001) [61]. Nowadays, Pareto distributions play an important role in the quantitative modelling of financial risk (see, e.g., Bouchaud and Potters (2003) [3]).

Working out the equation of the Lorenz curve associated with a Pareto distribution according to Eq. (7.29), using Eq. (8.73), yields a particularly simple result given by

$$L(\alpha;\gamma) = 1 - (1-\alpha)^{1-(1/\gamma)} \; . \tag{8.77}$$

This result forms the basis of Pareto's famous **80/20 rule** concerning concentration in the distribution of an asset of general importance in a given population. According to Pareto's empirical findings, typically 80% of such an asset are owned by just 20% of the regarded population (and vice versa) (cf. Pareto (1896) [39]).[3] The 80/20 rule applies exactly for a power-law index $\gamma = \dfrac{\ln(5)}{\ln(4)} \approx 1.16$. It is a prominent example of the phenomenon of **universality**, frequently observed in the mathematical modelling of quantitative–empirical relationships between variables in a wide variety of scientific disciplines; cf. Gleick (1987) [19, p 157ff].

In practice, bivariate quantitative–empirical data $\{(x_i, y_i)\}_{i=1,\ldots,n}$ for positive variables (X, Y) is tested for a Pareto distribution by investigating a scatter plot for the **logarithmic quantities** $\ln(y_i)$ against $\ln(x_i)$ for correlation; cf. Sec. 12.1. Given there is a functional relationship between Y and X of the form $y = Kx^{-(\gamma+1)}$, then the logarithmic quantities are related by

$$\ln(y) = \ln(K) - (\gamma+1) \times \ln(x) \; ,$$

i.e., there exists a *straight line relationship* between $\ln(y)$ and $\ln(x)$ with negative slope equal to $-(\gamma+1)$.

8.10 Power-law distribution

While the `pdf` of a Pareto distribution, discussed in the previous section, is proportional to positive powers of $1/x$, the slightly more general three-parameter **power-law distribution**,

$$X \sim Pl(a;b;c) \; . \tag{8.78}$$

includes cases with a `pdf` proportional to positive powers of x itself (i.e., for $c > 1$).

Spectrum of values:

$$X \mapsto x \in \{x | a \leq x \leq a+b, b \in \mathbb{R}_{>0}\} \subset \mathbb{R} \; . \tag{8.79}$$

Probability density function (`pdf`):

$$f_X(x) = \begin{cases} 0 & \text{for} \quad x < a \\ \dfrac{c}{b}\left(\dfrac{x-a}{b}\right)^{c-1}, \quad c \in \mathbb{R}_{>0} & \text{for} \quad a \leq x \leq a+b \\ 0 & \text{for} \quad x > b \end{cases} \; . \tag{8.80}$$

[3] See also footnote 2 in Subsec. 3.4.2.

Cumulative distribution function (`cdf`):

$$F_X(x) = P(X \leq x) = \begin{cases} 0 & \text{for} \quad x < a \\ \left(\dfrac{x-a}{b}\right)^c & \text{for} \quad a \leq x \leq a+b \\ 1 & \text{for} \quad x > b \end{cases} . \tag{8.81}$$

Expectation value and variance:

$$\mathrm{E}(X) = a + \frac{c}{c+1}\, b \tag{8.82}$$

$$\mathrm{Var}(X) = \frac{c}{(c+1)^2(c+2)}\, b^2 . \tag{8.83}$$

α–quantiles:

$$\alpha \stackrel{!}{=} F_X(x_\alpha) = \left(\frac{x_\alpha - a}{b}\right)^c \Leftrightarrow x_\alpha = F_X^{-1}(\alpha) = a + \sqrt[c]{\alpha} \times b \quad \text{for all} \quad 0 < \alpha < 1 . \tag{8.84}$$

8.11 Special hyperbolic distribution

The **special hyperbolic distribution** for a continuous random variable X,

$$X \sim sHyp , \tag{8.85}$$

which does not depend on any free parameters, can be considered an exotic but quite simple representative of a continuous probability distribution law.

Spectrum of values:

$$X \mapsto x \in [0,1] \subset \mathbb{R}_{\geq 0} . \tag{8.86}$$

Probability density function (`pdf`):

$$f_X(x) = \begin{cases} \dfrac{1}{\ln(2)}\, \dfrac{1}{1+x} & \text{for} \quad x \in [0,1] \\ 0 & \text{otherwise} \end{cases} ; \tag{8.87}$$

its graph is shown in Fig. 8.11 below.

Cumulative distribution function (`cdf`):

$$F_X(x) = P(X \leq x) = \begin{cases} 0 & \text{for} \quad x < 0 \\ \dfrac{1}{\ln(2)}\, \ln(1+x) & \text{for} \quad x \in [0,1] \\ 1 & \text{for} \quad x > 1 \end{cases} . \tag{8.88}$$

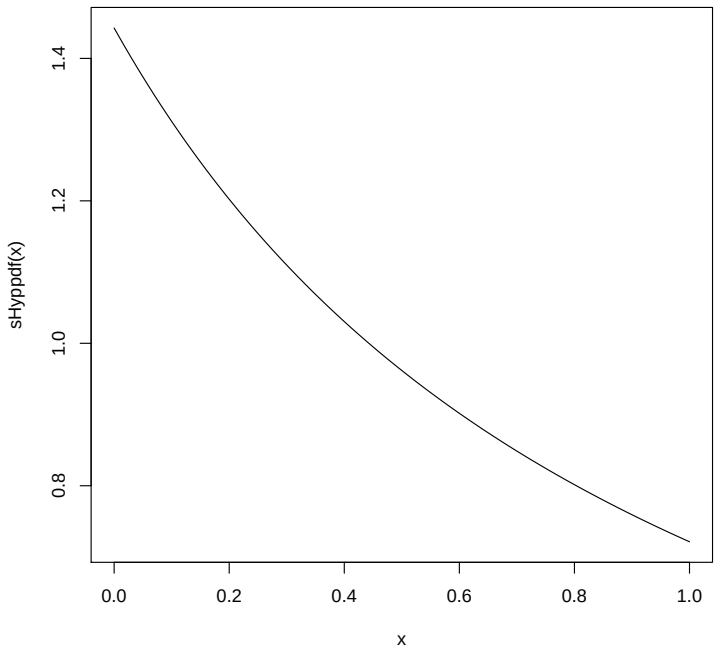

Figure 8.11: `pdf` of the special hyperbolic distribution according to Eq. (8.87).

Expectation value and variance:

$$\mathrm{E}(X) = \frac{1-\ln(2)}{\ln(2)} \tag{8.89}$$

$$\mathrm{Var}(X) = \frac{\frac{3}{2}\ln(2)-1}{(\ln(2))^2} \,. \tag{8.90}$$

α–quantiles:

$$\alpha \stackrel{!}{=} F_X(x_\alpha) = \frac{1}{\ln(2)}\,\ln(1+x_\alpha) \;\Leftrightarrow\; x_\alpha = F_X^{-1}(\alpha) = e^{\alpha\ln(2)} - 1 \quad \text{for all} \quad 0<\alpha<1\,. \tag{8.91}$$

8.12 Cauchy distribution

The French mathematician Augustin Louis Cauchy (1789–1857) is credited for the inception into **Statistics** of the continuous two-parameter distribution law

$$X \sim Ca(b;a)\,, \tag{8.92}$$

with properties

Spectrum of values:

$$X \mapsto x \in \mathbb{R}\,. \tag{8.93}$$

Probability density function (`pdf`):

$$\boxed{f_X(x) = \frac{1}{\pi}\,\frac{a}{a^2+(x-b)^2}\,, \qquad \text{with} \quad a \in \mathbb{R}_{>0},\ b \in \mathbb{R}\,;} \tag{8.94}$$

its graph is shown in Fig. 8.12 below for two particular cases.

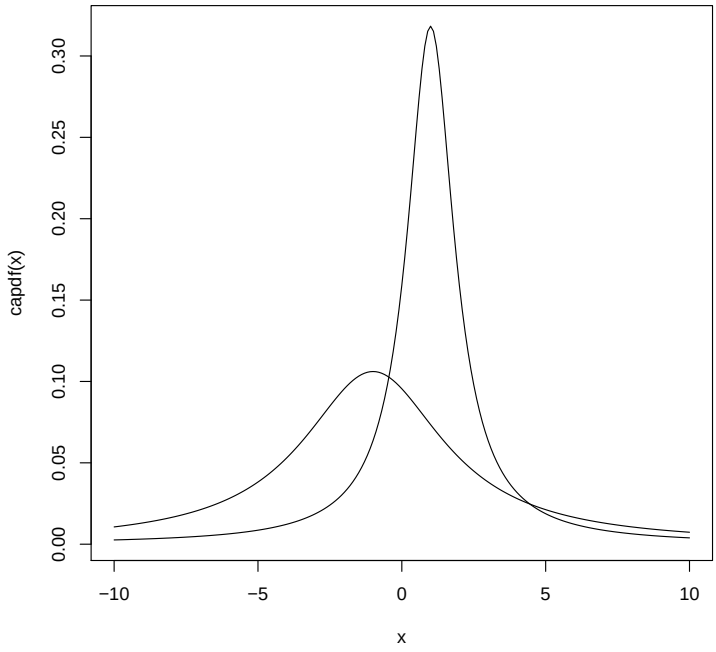

Figure 8.12: `pdf` of the Cauchy distribution according to Eq. (8.95). Displayed are the cases $Ca(1;1)$ (stongly peaked) and $Ca(-1;3)$ (moderately peaked).

Cumulative distribution function (`cdf`):

$$\boxed{F_X(x) = P(X \le x) = \frac{1}{2} + \frac{1}{\pi}\,\arctan\left(\frac{x-b}{a}\right)\,.} \tag{8.95}$$

Expectation value and variance:[4]

$$\mathrm{E}(X): \qquad \text{does NOT exist due to a diverging integral} \tag{8.96}$$

$$\mathrm{Var}(X): \qquad \text{does NOT exist due to a diverging integral}\,. \tag{8.97}$$

[4] In the case of a Cauchy distribution the fall-off in the tails of the `pdf` ist not sufficiently fast for the expectation value and variance integrals, Eqs. (7.26) and (7.27), to converge to finite values.

See, e.g., Sivia and Skilling (2006) [47, p 34].

α–quantiles:

$$\alpha \stackrel{!}{=} F_X(x_\alpha) \quad \Leftrightarrow \quad x_\alpha = F_X^{-1}(\alpha) = b + a \tan\left[\pi\left(\alpha - \frac{1}{2}\right)\right] \quad \text{for all} \quad 0 < \alpha < 1 \; . \tag{8.98}$$

8.13 Central limit theorem

Consider a set of n *mutually stochastically independent, additive* random variables $X_1, \ldots, X_n$ [cf. Eq. (6.13)], with (i) *finite* expectation values $\mu_1, \ldots, \mu_n$, (ii) *finite* variances $\sigma_1^2, \ldots, \sigma_n^2$ which are not too different from one another, and (iii) corresponding `cdf`s $F_1(x), \ldots, F_n(x)$. Define for this set a **total sum** Y_n and a **sample mean** $\bar{X}_n$ according to Eq. (7.34). In analogy to Eq. (7.32), a **standardised summation random variable** associated with Y_n is given by

$$Z_n := \frac{Y_n - \sum_{i=1}^n \mu_i}{\sqrt{\sum_{j=1}^n \sigma_j^2}} \; . \tag{8.99}$$

Subject to the convergence condition

$$\lim_{n\to\infty} \max_{1\leq i\leq n} \frac{\sigma_i}{\sqrt{\sum_{j=1}^n \sigma_j^2}} = 0 \; , \tag{8.100}$$

i.e., that asymptotically the standard deviation of the total sum dominates the standard deviations of all individual X_i, and certain additional regularity requirements (cf. Rinne (2008) [45, p 427 f]), the **central limit theorem** in its general form according to the Finnish mathematician Jarl Waldemar Lindeberg (1876–1932) and the Croatian–American mathematician William Feller (1906–1970) states that in the asymptotic limit of infinitely many X_i,

$$\lim_{n\to\infty} F(z_n) = \Phi(z) \; , \tag{8.101}$$

i.e., the limit distribution of the standardised summation random variable Z_n is given by the **standard normal distribution** $N(0;1)$ introduced in Sec. 8.5; cf. Lindeberg (1922) [32] and Feller (1951) [13]. Earlier results on the asymptotic distributional properties of a sum of independent random variables were obtained by the Russian mathematician, mechanician and physicist Aleksandr Mikhailovich Lyapunov (1857–1918); cf. Lyapunov (1901) [34]. Thus, under fairly general conditions, the normal distribution acts as an **attractor distribution** for the sum of n stochastically independent, additive random variables X_i.[5] In oversimplified terms: this result bears a certain economical convenience for most practical purposes in that, given favourable conditions, when the size of a random sample is sufficiently large (in practice a typical rule of thumb is $n \geq 50$), one essentially needs to know the characteristic features of only a single continuous

[5] Put differently, for increasingly large n the `cdf` of the total sum Y_n approximates a normal distribution with expectation value $\sum_{i=1}^n \mu_i$ and variance $\sum_{i=1}^n \sigma_i^2$ to an increasingly accurate degree. In particular, all reproductive distributions may be approximated by a normal distribution as n becomes large.

probability distribution law to perform, e.g., the statistical testing of hypotheses; cf. Ch. 10. As will become apparent in subsequent chapters, the central limit theorem has profound ramifications for applications in all empirical scientific disciplines.

Note that for *finite* n the central limit theorem makes *no* statement as to the nature of the *tails* of the distributions of Z_n and of Y_n (where, in principle, it can be very different from a normal distribution; cf. Bouchaud and Potters (2003) [3, p 25f]).

A direct consequence of the central limit theorem and its preconditions is the fact that for the **sample mean** $\bar{X}_n$ both

$$\lim_{n\to\infty} \mathrm{E}(\bar{X}_n) = \lim_{n\to\infty} \frac{\sum_{i=1}^{n} \mu_i}{n} \qquad \text{and} \qquad \lim_{n\to\infty} \mathrm{Var}(\bar{X}_n) = \lim_{n\to\infty} \frac{\sum_{i=1}^{n} \sigma_i^2}{n^2}$$

converge to finite values. This property is most easily recognised in the special case of n **stochastically independent and identically distributed** (in short: "i.i.d") additive random variables $X_1, \ldots, X_n$, with common finite expectation value μ, finite variance σ^2, and `cdf` $F(x)$.[6] Then,

$$\lim_{n\to\infty} \mathrm{E}(\bar{X}_n) \;=\; \lim_{n\to\infty} \frac{n\mu}{n} \;=\; \mu \tag{8.102}$$

$$\lim_{n\to\infty} \mathrm{Var}(\bar{X}_n) \;=\; \lim_{n\to\infty} \frac{n\sigma^2}{n^2} \;=\; \lim_{n\to\infty} \frac{\sigma^2}{n} \;=\; 0\,, \tag{8.103}$$

which is known as the **law of large numbers** according to the Swiss mathematician Jacob Bernoulli (1654–1705); the sample mean $\bar{X}_n$ **converges stochastically** to its expectation value μ.

This ends Part II of these lecture notes and we now turn to Part III in which we focus on a number of useful applications of **inferential statistical methods** of **data analysis**.

[6] These conditions lead to the central limit theorem in the special form according to Jarl Waldemar Lindeberg (1876–1932) and the French mathematician Paul Pierre Lévy (1886–1971).

Chapter 9

Operationalisation of latent variables: Likert's scaling method of summated item ratings

The most frequently practiced method to date of operationalising **latent variables** (such as "social constructs") in the **Social Sciences** and **Humanities** is due to the US-American psychologist Rensis Likert's (1903–1981). In his 1932 paper [31], which completed his thesis work for a Ph.D., he expressed the idea that **latent variables** X_L, when they are perceived of as *one-dimensional* in nature, can be rendered measurable in a *quasi-metrical* fashion by means of the **summated ratings** over an extended list of suitable **indicator items** X_i ($i = 1, 2, \ldots$). Such indicator items are often formulated as specific statements relating to the theoretical concepts underlying a particular 1–D latent variable X_L to be measured, with respect to which test persons need to express their level of agreement. A typical **response scale** for the items X_i, providing the necessary item ratings, is given for instance by the 5–level ordinally ranked attributes of agreement

1: strongly disagree/strongly unfavourable

2: disagree/unfavourable

3: undecided

4: agree/favourable

5: strongly agree/strongly favourable.

In the research literature, one also encounters 7–level or 10–level item rating scales. Note that it is *assumed* that the items X_i, and thus their ratings, can be treated as **additive**, so that the conceptual principles of Sec. 7.5 relating to sums of random variables can be relied upon. When forming the sum over the ratings of all indicator items selected, it is essential to carefully pay attention to the **polarity** of the items involved. For the resultant **total sum** $\sum_i X_i$ to be consistent, the polarity of all items used needs to be uniform.[1]

[1] For a questionnaire, however, it is strongly recommended to include also indicator items of reversed polarity. This will improve the overall construct validity of the measurement tool.

The construction of a consistent **Likert scale** for a 1–D latent variable X_L involves four basic steps (see, e.g., Trochim (2006) [58]):

(i) the compilation of an initial list of 80 to 100 potential **indicator items** X_i for the latent variable of interest,

(ii) the draw of a **gauge random sample** from the targeted population $\mathbf{\Omega}$,

(iii) the computation of the **total sum** $\sum_i X_i$ of item ratings, and, most importantly,

(iv) the performance of an **item analysis** based on the sample data and the associated total sum $\sum_i X_i$ of item ratings.

The item analysis, in particular, consists of the consequential application of two exclusion criteria. Items are being discarded from the list when either

(a) they show a weak **item-to-total correlation** with the total sum $\sum_i X_i$ (a rule of thumb is to exclude items with correlations less than 0.5), or

(b) it is possible to increase the value of **Cronbach's**[2] $\boldsymbol{\alpha}$**–coefficient** (see Cronbach (1951) [9]), a measure of the scale's **internal consistency reliability**, by excluding a particular item from the list (the objective being to attain α-values greater than 0.8).

For a set of $m \in \mathbb{N}$ indicator items X_i, Cronbach's α–coefficient is defined by

$$\boxed{\alpha := \left(\frac{m}{m-1}\right)\left(1 - \frac{\sum_{i=1}^{m} S_i^2}{S_{\text{total}}^2}\right)\ ,} \tag{9.1}$$

where S_i^2 denotes the sample variance associated with the ith indicator item, and S_{total}^2 is the sample variance of the total sum $\sum_i X_i$.

SPSS: Analyze → Scale → Reliability Analysis ... (Model: Alpha) → Statistics ...: Scale if item deleted
R: `alpha(`*items*`)` (package: `psych`)

The outcome of the item analysis is a drastic reduction of the initial list to a set of $k \in \mathbb{N}$ items X_i ($i = 1, \ldots, k$) of high discriminatory power (where typically k is an integer in the range of 10 to 15). The associated **total sum**

$$X_L := \sum_{i=1}^{k} X_i \tag{9.2}$$

thus operationalises the 1–D latent variable X_L in a quasi-metrical fashion since it is to be measured on an **interval scale** with a discretised spectrum of values given by

$$X_L \mapsto \sum_{i=1}^{k} x_i \in [1k, 5k]\ . \tag{9.3}$$

1–D latent variable X_L:

• **Item** X_1:	strongly disagree	◯	◯	◯	◯	◯	strongly agree
• **Item** X_2:	strongly disagree	◯	◯	◯	◯	◯	strongly agree
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
• **Item** X_k:	strongly disagree	◯	◯	◯	◯	◯	strongly agree

Table 9.1: Structure of a k-indicator-item Likert scale for some 1–D latent variable X_L, based on a visualised equidistant 5–level item rating scale.

The structure of a finalised k-indicator-item Likert scale for some 1–D latent variable X_L with an equidistant graphical 5–level item rating scale is displayed in Tab. 9.1.

Likert's scaling method of aggregating information from a set of k ordinally scaled items to form an effectively quasi-metrical total sum $X_L = \sum_i X_i$ draws its legitimisation to a large extent from the **central limit theorem** (cf. Sec. 8.13). In practice it is found that in many cases of interest the total sum $X_L = \sum_i X_i$ is normally distributed in its samples to a very good approximation. The main shortcoming of Likert's approach is its dependency of the gauging process of the scale on the targeted population.

[2]Named after the US-American educational psychologist Lee Joseph Cronbach (1916–2001). The range of the normalised α–coefficient is the real-valued interval $[0, 1]$.

Chapter 10

Random sampling of populations and statistical testing of hypotheses

Quantitative–empirical research methods may be employed for **exploratory** as well as for **confirmatory data analysis**. Here we will focus on the latter. To investigate **research questions** systematically by statistical means, with the objective to make inferences about the distributional properties of a set of **statistical variables** on the basis of analysis of data from just a few units in a sample $\boldsymbol{S}_{\boldsymbol{\Omega}}$ to an entire population $\boldsymbol{\Omega}$ worth of statistical units, the following three issues have to be addressed in a clearcut fashion:

(i) the target **population** $\boldsymbol{\Omega}$ of the research activity needs to be defined in an unambiguous way,

(ii) an adequate **random sample** $\boldsymbol{S}_{\boldsymbol{\Omega}}$ needs to be drawn from $\boldsymbol{\Omega}$, and

(iii) a reliable mathematical procedure for **estimating quantitative population parameters** from random sample data needs to be employed.

We will briefly discuss these issues in turn, beginning with a review in Tab. 10.1 of conventional **notation** for distinguishing specific statistical measures relating to populations from the corresponding ones relating to random samples. Towards the end of this chapter, we will highlight the principles underlying a systematic **statistical testing of hypotheses** regarding specific distributional features of observable quantities.

Random variables in a population $\boldsymbol{\Omega}$ (of size N) will be denoted by capital Latin letters such as X, Y, ..., Z, while their **realisations** in random samples $\boldsymbol{S}_{\boldsymbol{\Omega}}$ (of size n) will be denoted by lower case Latin letters such as x_i, y_i, ..., z_i ($i = 1, \ldots, n$). In addition, one denotes **population parameters** by lower case Greek letters, while for their corresponding **point estimator functions** relating to random samples, which are also perceived of as random variables, again capital Latin letters are used for representation. The ratio n/N will be referred to as the **sampling fraction**. As is standard in the literature, we will denote a particular **random sample** of size n for a random variable X by a set $\boldsymbol{S}_{\boldsymbol{\Omega}}$: $(X_1, \ldots, X_n)$.

In actual practice, it is often not possible to acquire access for the purpose of enquiry to every single statistical unit belonging to an identified target population $\boldsymbol{\Omega}$, not even in principle. For example, this could be due to the fact that $\boldsymbol{\Omega}$'s size N is far too large to be determined accurately.

Population Ω	**Random sample $\boldsymbol{S_\Omega}$**
population size N	sample size n
arithmetical mean μ	sample mean $\bar{X}_n$
standard deviation σ	sample standard deviation S_n
median $\tilde{x}_{0.5}$	sample median $\tilde{X}_{0.5,n}$
correlation coefficient ρ	sample correlation coefficient r
rank correlation coefficient ρ_S	sample rank correl. coefficient r_S
regression coefficient (intercept) α	sample regression intercept a
regression coefficient (slope) β	sample regression slope b

Table 10.1: Notation for distinguishing between statistical measures relating to a population $\boldsymbol{\Omega}$ on the one-hand side, and to the corresponding quantities and unbiased point estimator functions obtained from a random sample $\boldsymbol{S_\Omega}$ on the other.

In this case, to ensure a thorough investigation, one needs to resort to using a **sampling frame** representative of $\mathbf{\Omega}$. By this one understands a representative list of elements in $\mathbf{\Omega}$ to which access may actually be obtained. Such a list will have to be compiled by some authority of scientific integrity. In an attempt to avoid a notational overflow, we will subsequently continue to use N to denote *both*: the size of $\mathbf{\Omega}$ and the size of its associated **sampling frame** (even though this is not entirely correct).

We now proceed to introduce the three most commonly practiced methods of drawing **random samples** from given fixed populations $\mathbf{\Omega}$ of statistical units.

10.1 Random sampling methods

10.1.1 Simple random sampling

The **simple random sampling** technique can be best understood in terms of the **urn model** of **combinatorics** introduced in Sec. 6.4. Given a population $\mathbf{\Omega}$ of N distinguishable statistical units, there is a total of $\left(\begin{array}{c} N \\ n \end{array} \right)$ distinct possibilities of drawing samples of size n from $\mathbf{\Omega}$, given the order of selection is *not* being accounted for. A **simple random sample** is then defined by the property that its probability of selection is equal to

$$\frac{1}{\left(\begin{array}{c} N \\ n \end{array} \right)} \, , \tag{10.1}$$

according to the Laplacian principle of Eq. (6.8). This has the immediate consequence that the *a priori* probability of selection of any single statistical unit is given by[1]

$$1 - \frac{\left(\begin{array}{c} N-1 \\ n \end{array} \right)}{\left(\begin{array}{c} N \\ n \end{array} \right)} = 1 - \frac{N-n}{N} = \frac{n}{N} \, . \tag{10.2}$$

On the other hand, the probability that two statistical units i and j are being selected for the same sample amounts to

$$\frac{n}{N} \times \frac{n-1}{N-1} \, . \tag{10.3}$$

As such, by Eq. (6.13), the selection of two statistical units is *not* stochastically independent (in which case the joint probability would be $n/N \times n/N$). However, for sampling fractions $n/N \leq 0.05$, stochastic independence of the selection of statistical units generally holds to a reasonably good approximation. When, in addition, $n \geq 50$, likewise the conditions for the **central limit theorem** in the variant of Lindeberg and Lévy to apply (cf. Sec. 8.13) hold to a rather good degree.

[1] In the literature this particular property of a random sample is referred to as "epsem": equal probability of selection method.

10.1.2 Stratified random sampling

Stratified random sampling adapts the sampling process to an intrinsic structure of the population $\mathbf{\Omega}$ as provided by the k mutually exclusive and exhaustive categories of some qualitative (nominal or ordinal) variable, which thus defines a set of k **strata** (layers) of $\mathbf{\Omega}$. By construction, there are N_i statistical units belonging to the ith stratum ($i = 1, \ldots, k$). Simple random samples of sizes n_i are drawn from each stratum according to the principles outlined in Subsec. 10.1.1, yielding a total sample of size $n = n_1 + \ldots + n_k$. Frequently applied variants of this sampling technique are (i) **proportionate allocation** of statistical units, defined by the condition[2]

$$\frac{n_i}{n} \stackrel{!}{=} \frac{N_i}{N} \qquad \Rightarrow \qquad \frac{n_i}{N_i} = \frac{n}{N} \; ; \tag{10.4}$$

in particular, this allows for a fair representation of minorities in $\mathbf{\Omega}$, and (ii) **optimal allocation** of statistical units which aims at a minimisation of the resultant sample variances of the variables investigated. Further details on the stratified random sampling technique can be found, e.g., in Bortz and Döring (2006) [5, p 425ff].

10.1.3 Cluster random sampling

When the population $\mathbf{\Omega}$ naturally subdivides into an exhaustive set of K mutually exclusive **clusters** of statistical units, a convenient sampling strategy is given by selecting $k < K$ clusters from the set at random and perform complete surveys within each of the chosen clusters. The probability of selection of any particular statistical unit from $\mathbf{\Omega}$ thus amounts to k/K. This **cluster random sampling** method has the practical advantage of being less contrived. However, in general it entails sampling errors that are greater than in the previous two sampling approaches. Further details on the cluster random sampling technique can be found, e.g., in Bortz and Döring (2006) [5, p 435ff].

10.2 Point estimator functions

Many inferential statistical methods of data analysis revolve around the **estimation** of unknown **distribution parameters** θ with respect to some population $\mathbf{\Omega}$ by means of corresponding **point estimator functions** $\hat{\theta}_n(X_1, \ldots, X_n)$ (or **statistics**), the values of which are computed from the data of a **random sample** $\boldsymbol{S}_{\mathbf{\Omega}}$: $(X_1, \ldots, X_n)$. Owing to the stochastic nature of the random sampling approach, any point estimator function $\hat{\theta}_n(X_1, \ldots, X_n)$ is subject to a **random sampling error**. One can show that this estimation procedure becomes reliable provided that a point estimator function satisfies the following two important criteria of quality:

(i) **Unbiasedness:** $\mathrm{E}(\hat{\theta}_n) = \theta$

(ii) **Consistency:** $\lim\limits_{n \to \infty} \mathrm{Var}(\hat{\theta}_n) = 0$.

[2] Note that, thus, this also has the "epsem" property.

For metrically scaled random variables X, defining for a given random sample $\boldsymbol{S_\Omega}$: $(X_1, \ldots, X_n)$ of size n a **sample total** by

$$Y_n := \sum_{i=1}^{n} X_i \ , \tag{10.5}$$

the two most prominent **point estimator functions** satisfying the **unbiasedness** and **consistency** conditions are the

$$\textbf{sample mean}: \qquad \bar{X}_n := \frac{1}{n} Y_n \tag{10.6}$$

$$\textbf{sample variance}: \qquad S_n^2 := \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X}_n)^2 \ . \tag{10.7}$$

These will be frequently employed in subsequent considerations for point-estimating the population parameters μ and σ^2. **Sampling theory** holds it that the sizes of the **standard errors** associated with the point estimator functions (10.6) and (10.7) amount to the standard deviations of the underlying theoretical **sampling distributions** of these functions. For given population $\boldsymbol{\Omega}$ of size N, imagine drawing a very long sequence of mutually independent random samples of a fixed size n, from each of which individual realisations of $\bar{X}_n$ and S_n^2 are obtained. In the limit that this sequence becomes infinitely long, and n is kept fixed, the theoretical **sampling distributions** of $\bar{X}_n$ and S_n^2 are normal (cf. Sec. 8.5) resp. χ^2 with $n-1$ degrees of freedom (cf. Sec. 8.6), with standard deviations

$$\frac{S_n}{\sqrt{n}} \qquad \text{resp.} \qquad \sqrt{\frac{2}{n-1}}\, S_n^2 \ ; \tag{10.8}$$

cf., e.g., Lehman and Casella (1998) [28, p 91ff], and Levin *et al* (2010) [30, Ch. 6]. Thus, for a *finite* sample standard deviation S_n, these two **standard errors** decrease with the sample size n proportional to the inverse of $\sqrt{n}$ resp. the inverse of $\sqrt{n-1}$.

10.3 Statistical tests of hypotheses

10.3.1 General procedure

The **statistical testing of hypotheses** by means of observable quantities is the centrepiece of the current body of inferential statistical methods. Its logic of an ongoing routine of a systematic **falsification** of hypotheses by empirical means is firmly rooted in the ideas of **critical rationalism** and **logical positivism**, as expressed most emphatically by the Austro–British philosopher Sir Karl Raimund Popper CH FRS FBA (1902–1994); see, e.g., Popper (2002) [44]. The systematic procedure for statistically testing hypotheses, as practiced today as a standardised method of probability-based decision-making, was developed during the first half of the 20^{th} Century, predominantly by the English statistician, evolutionary biologist, eugenicist and geneticist Sir Ronald Aylmer Fisher FRS (1890–1962), the Polish–US-American mathematician and statistician Jerzy Neyman (1894–1981), the English mathematician and statistician Karl Pearson FRS (1857–1936), and his son, the English statistician Egon Sharpe Pearson CBE FRS (1895–1980); cf. Fisher (1935) [16], Neyman and Pearson

(1933) [38], and Pearson (1900) [40]. We will describe the main steps of the systematic test procedure in the following.

The central aim of the statistical testing of hypotheses is to separate **true effects** in a targeted **population** Ω of statistical units concerning distributional properties of, or relations between, selected **statistical variables** $X, Y, \ldots, Z$ from **chance effects** due to the sampling approach to probing the nature of Ω. The latter results in a generally unavoidable state of incomplete information on the part of the researcher.

In an inferential statistical context, **hypotheses** are formulated as assumptions on

(i) the **probability distribution function** F of one or more **random variables** $X, Y, \ldots, Z$ in Ω, or

(ii) one or more **parameters** θ of this distribution function.

Generically, statistical hypotheses need to be viewed as probabilistic statements. As such the researcher will always have to deal with a fair amount of **uncertainty** in deciding whether a particular effect is **significant** in Ω or not. Bernstein (1998) [2, p 207] summarises the circumstances relating to the test of a specific hypothesis as follows:

> "Under conditions of uncertainty, the choice is not between rejecting a hypothesis and accepting it, but between reject and not–reject."

The question arises as to *which kinds of quantitative problems can be efficiently settled by statistical means*? With respect to a given target population Ω, in the simplest kinds of applications of hypothesis tests, one may (a) **test for differences** in the distributional properties of a single statistical variable X between a number of subgroups of Ω, necessitating **univariate methods** of data analysis, or one may (b) **test for association** between two statistical variables X and Y, thus requiring **bivariate methods** of data analysis. The standardised test procedure takes the following steps on the way to making a decision:

Schematic of the statistical test of a hypothesis

1. Formulation, with respect to Ω, of a pair of mutually exclusive **hypotheses**:

   (a) the **null hypothesis** H_0 conjectures that "there exists *no* effect in Ω of the kind envisaged," while

   (b) the **research hypothesis** H_1 conjectures that "there *does* exist a true effect in Ω of the kind envisaged."

   The starting point of the procedure is the *assumption (!)* that it is the H_0 conjecture which holds true in Ω. The objective is to try to refute H_0 empirically on the basis of random sample data drawn from Ω, to a certain level of significance which needs to be specified in advance. In this sense it is H_0 which is being subjected to a statistical test.[3] The striking *asymmetry* regarding the roles of H_0 and H_1 in the test procedure forms the basis of the method of **falsification** of hypotheses advocated by critical rationalism.

[3] Bernstein (1998) [2, p 209] refers to the statistical test of a hypothesis as a "mathematical stress test".

2. Fixing of a **significance level** α prior to the test, where, by convention, $\alpha \in [0.01, 0.05]$. The parameter α is synonymous with the probability of committing a Type I error (to be defined below) in making a test decision.

3. Construction of a suitable continuous real-valued measure, a **test statistic** $T_n(X_1, \ldots, X_n)$, for quantifying deviations of the data from a random sample $\boldsymbol{S_\Omega}$: $(X_1, \ldots, X_n)$ of size n from the initial "no effect in $\boldsymbol{\Omega}$" conjecture of H_0, with *known (!)* associated **theoretical distribution** for computing corresponding event probabilities.

4. Determination of the **rejection region** B_α for H_0 within the spectrum of values of the test statistic $T_n(X_1, \ldots, X_n)$ from re-arranging the condition

$$P\left(T_n(X_1, \ldots, X_n) \in B_\alpha | H_0\right) \stackrel{!}{\leq} \alpha \, . \tag{10.9}$$

5. Computation of a specific **realisation** $t_n(x_1, \ldots, x_n)$ of the test statistic $T_n(X_1, \ldots, X_n)$ from data $x_1, \ldots, x_n$ of a **random sample $\boldsymbol{S_\Omega}$:** $(X_1, \ldots, X_n)$.

6. Obtaining a **test decision** on the basis of one of the alternative criteria: when for the realisation $t_n(x_1, \ldots, x_n)$ of the test statistic $T_n(X_1, \ldots, X_n)$, resp. the corresponding p–value (to be defined in Subsec. 10.3.2 below),[4]

   (i) $t_n \in B_\alpha$, resp. $\boldsymbol{p}$**–value** $< \alpha \quad \Rightarrow \quad$ reject H_0,

   (ii) $t_n \notin B_\alpha$, resp. $\boldsymbol{p}$**–value** $\geq \alpha \quad \Rightarrow \quad$ not reject H_0.

When performing a statistical test of a **null hypothesis** H_0, the researcher is in **risk** of making a wrong decision. Hereby, one distinguishes the following two kinds of error:

Type I error: reject an H_0 which, however, is true, with probability $P(H_1|H_0 \text{ true}) = \alpha$, and
Type II error: not reject an H_0 which, however, is false, with probability $P(H_0|H_1 \text{ true}) = \beta$.

By fixing the significance level α prior to running a statistical test, one simultaneously controls the risk of a Type I error. We condense the different possible outcomes when making a test decision in the following table:

[4]The statistical software packages SPSS and R provide p–values as a means for making decisions in the statistical testing of hypotheses.

Consequences of test decisions:

	H_0	Decision for:	H_1
H_0 true	correct decision: $P(H_0\|H_0 \text{ true}) = 1-\alpha$		**Type I error**: $P(H_1\|H_0 \text{ true}) = \alpha$
Reality / Ω:			
H_1 true	**Type II error**: $P(H_0\|H_1 \text{ true}) = \beta$		correct decision: $P(H_1\|H_1 \text{ true}) = 1-\beta$

While the probability α is required to be specified *a priori* to a statistical test, the probability β can be computed only *a posteriori*. One refers to the probability $1-\beta$ associated with the latter as the **power** of a particular statistical test.

Note that in the complementary **Bayesian approach** to **statistical data analysis** (cf. Subsec. 6.5.2) the empirical testing of hypotheses follows the logic

$$P(\text{hypothesis}|\text{data}) \propto P(\text{data}|\text{hypothesis}) \times P(\text{hypothesis}) \,, \tag{10.10}$$

thus requiring information on (i) the **joint probability distributions** of parameters and random variables (with parameters treated as random variables) and the distribution of random sample data underlying the probability $P(\text{data}|\text{hypothesis})$, as well as specification of (ii) a *subjective* prior probability $P(\text{hypothesis})$. This information is not necessarily always available though. Further details on the Bayesian method are given in, e.g., Sivia and Skilling (2006) [47, p 6], or Lupton (1993) [33, p 50ff].

10.3.2 Definition of a p–value

Def.: Let $T_n(X_1,\ldots,X_n)$ be the **test statistic** underlying a particular **statistical test**, the **theoretical distribution** of which be *known* under the assumption that the null hypothesis H_0 holds true in Ω. If $t_n(x_1,\ldots,x_n)$ is the **realisation** of $T_n(X_1,\ldots,X_n)$ computed from the data $x_1,\ldots,x_n$ of a random sample $\boldsymbol{S}_{\boldsymbol{\Omega}}$: $(X_1,\ldots,X_n)$, then the ***p*–value** associated with $t_n(x_1,\ldots,x_n)$ is defined as the probability of obtaining a value of $T_n(X_1,\ldots,X_n)$ which is *more extreme* than the given $t_n(x_1,\ldots,x_n)$, given that the null hypothesis applies.

Specifically, using the computational rules (7.22)–(7.24), one obtains for a

- two-sided statistical test,

$$\begin{aligned} p &:= P(T_n < -|t_n|\,|\,H_0) + P(T_n > |t_n|\,|\,H_0) \\ &= P(T_n < -|t_n|\,|\,H_0) + 1 - P(T_n \leq |t_n|\,|\,H_0) \\ &= F_{T_n}(-|t_n|) + 1 - F_{T_n}(|t_n|) \,. \end{aligned} \tag{10.11}$$

This result specialises to $p = 2\,[1 - F_{T_n}(|t_n|)]$ if the respective pdf exhibits **reflection symmetry** with respect to a vertical axis at $t_n = 0$, i.e., when $F_{T_n}(-|t_n|) = 1 - F_{T_n}(|t_n|)$ holds.

- left-sided statistical test,

$$p := P(T_n < t_n|H_0) = F_{T_n}(t_n)\ , \tag{10.12}$$

- right-sided statistical test,

$$p := P(T_n > t_n|H_0) = 1 - P(T_n \leq t_n|H_0) = 1 - F_{T_n}(t_n)\ . \tag{10.13}$$

With respect to the **test decision criterion** of rejecting an H_0 whenever $p < \alpha$, one refers to (i) cases with $p < 0.05$ as **significant** test results, and to (ii) cases with $p < 0.01$ as **highly significant** test results.

Remark: User-friendly routines for the computation of p–values are available in SPSS, R and EXCEL, and also on some GDCs.

In the following two chapters, we will turn to discuss a number of standard problems in **Inferential Statistics**, in association with the quantitative–empirical tools that have been developed to tackle them. In Ch. 11 we will be concerned with problems of a **univariate** nature, in particular, **testing for differences** in the distributional properties of a single random variable X between two of more subgroups of some population $\mathbf{\Omega}$, while in Ch. 12 the problems at hand will be of a **bivariate** nature, **testing for statistical association** in $\mathbf{\Omega}$ between pairs of random variables (X, Y).

Chapter 11

Univariate methods of statistical data analysis: confidence intervals and testing for differences

In this chapter we present a selection of standard inferential statistical techniques that, based on random sampling of some population $\mathbf{\Omega}$, were developed for the purpose of (a) estimating unknown distribution parameters by means of **confidence intervals**, (b) **testing for differences** between a given empirical distribution of a random variable and its *a priori* assumed theoretical distribution, and (c) **comparing** distributional properties and parameters of a random variable between two or more subgroups of $\mathbf{\Omega}$. Since the methods to be introduced relate to considerations on distributions of a single random variable only, they are thus referred to as **univariate**.

11.1 Confidence intervals

Assume given a continuous random variable X which satisfies in some population $\mathbf{\Omega}$ a **Gaußian normal distribution** with unknown distribution parameters μ and σ^2 (cf. Sec. 8.5). The issue is to determine, using data from a random sample $\boldsymbol{S}_{\mathbf{\Omega}}$: $(X_1, \ldots, X_n)$, a two-sided **confidence interval** estimate for any one of these unknown **distribution parameters** θ which can be relied on at a **confidence level** $1 - \alpha$, where, by convention, $\alpha \in [0.01, 0.05]$. Centred on a suitable unbiased and consistent point estimator function $\hat{\theta}_n(X_1, \ldots, X_n)$, the aim of the estimation process is to explicitly account for the **sampling error** δ_K arising due to the random selection process. This approach yields a two-sided confidence interval

$$K_{1-\alpha}(\theta) = \left[\hat{\theta}_n - \delta_K, \hat{\theta}_n + \delta_K\right] \ , \tag{11.1}$$

such that $P(\theta \in K_{1-\alpha}(\theta)) = 1 - \alpha$ applies. In the following we will consider the two cases which result when choosing $\theta \in \{\mu, \sigma^2\}$.

11.1.1 Confidence intervals for a population mean

When $\theta = \mu$, and $\hat{\theta}_n = \bar{X}_n$ by Eq. (10.6), the **two-sided confidence interval for a population mean** μ at significance level $1 - \alpha$ becomes

$$K_{1-\alpha}(\mu) = \left[\bar{X}_n - \delta_K, \bar{X}_n + \delta_K\right] \ , \tag{11.2}$$

with a **sampling error** amounting to

$$\delta_K = t_{n-1;1-\alpha/2} \, \frac{S_n}{\sqrt{n}} \ , \tag{11.3}$$

where S_n is the positive square root of the **sample variance** S_n^2 according to Eq. (10.7), and $t_{n-1;1-\alpha/2}$ denotes the value of the $(1-\alpha/2)$–quantile of a t–distribution with $df = n - 1$ degrees of freedom; cf. Sec. 8.7. The ratio $\dfrac{S_n}{\sqrt{n}}$ is the **standard error** associated with $\bar{X}_n$.

GDC: mode `STAT` $\rightarrow$ `TESTS` $\rightarrow$ `TInterval`

Equation (11.3) may be inverted to obtain the **minimum sample size** necessary to construct a two-sided confidence interval for μ to a prescribed accuracy $\delta_{\max}$, maximal sample variance $\sigma^2_{\max}$, and at fixed confidence level $1 - \alpha$. Thus,

$$n \geq \left(\frac{t_{n-1;1-\alpha/2}}{\delta_{\max}}\right)^2 \sigma^2_{\max} \ . \tag{11.4}$$

11.1.2 Confidence intervals for a population variance

When $\theta = \sigma^2$, and $\hat{\theta}_n = S_n^2$ by Eq. (10.7), the associated point estimator function

$$\frac{(n-1)S_n^2}{\sigma^2} \sim \chi^2(n-1) \ , \quad \text{with} \quad n \in \mathbb{N} \ , \tag{11.5}$$

satisfies a χ^2–distribution with $df = n - 1$ degrees of freedom; cf. Sec. 8.6. By inverting the condition

$$P\left(\chi^2_{n-1;\alpha/2} \leq \frac{(n-1)S_n^2}{\sigma^2} \leq \chi^2_{n-1;1-\alpha/2}\right) \stackrel{!}{=} 1 - \alpha \ , \tag{11.6}$$

one derives a **two-sided confidence interval for a population variance** σ^2 at significance level $1 - \alpha$ given by

$$\left[\frac{(n-1)S_n^2}{\chi^2_{n-1;1-\alpha/2}}, \frac{(n-1)S_n^2}{\chi^2_{n-1;\alpha/2}}\right] \ . \tag{11.7}$$

$\chi^2_{n-1;\alpha/2}$ and $\chi^2_{n-1;1-\alpha/2}$ again denote the values of particular quantiles of a χ^2–distribution.

11.2 One-sample χ^2–goodness–of–fit–test

A standard research question in quantitative–empirical investigations deals with the issue whether or not, with respect to some population $\mathbf{\Omega}$ of sample units, the **distribution law** of a specific random variable X may be assumed to comply with a particular theoretical reference distribution. This question can be formulated in terms of the corresponding `cdf`s, $F_X(x)$ and $F_0(x)$, presupposing that for practical reasons the spectrum of values of X is subdivided into a set of k mutually exclusive **categories** (or bins), with k a judiciously chosen integer which depends in the first place on the size n of the random sample $\boldsymbol{S_\Omega}$: $(X_1, \ldots, X_n)$ to be investigated.

The non-parametric **one-sample χ^2–goodness–of–fit–test** takes as its starting point the pair of

Hypotheses:

$$\begin{cases} H_0 : F_X(x) = F_0(x) & \Leftrightarrow & O_i - E_i = 0 \\ H_1 : F_X(x) \neq F_0(x) & \Leftrightarrow & O_i - E_i \neq 0 \end{cases} , \tag{11.8}$$

where O_i ($i = 1, \ldots, k$) denotes the actually **observed frequency** of category i in a random sample of size n, $E_i := np_i$ denotes the under H_0 (and so $F_0(x)$) theoretically **expected frequency** of category i in the same random sample, and p_i is the **probability** of finding a value of X in category i under $F_0(x)$.

The present procedure, devised by Pearson (1900) [40], employs the **residuals** $O_i - E_i$ ($i = 1 \ldots, k$) to construct a suitable

Test statistic:

$$\boxed{T_n(X_1, \ldots, X_n) = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i} \overset{H_0}{\approx} \chi^2(k-1-r)} \tag{11.9}$$

in terms of a sum of rescaled squared residuals $\dfrac{(O_i - E_i)^2}{E_i}$, which, under H_0, approximately satisfies a χ^2–distribution with $df = k - 1 - r$ degrees of freedom (cf. Sec. 8.6), where r is the number of unknown parameters of the reference distribution $F_0(x)$ which need to be estimated from the random sample data. For this test procedure to be reliable it is *important (!)* that the size n of the random sample be chosen such that the condition

$$E_i \overset{!}{\geq} 5 \tag{11.10}$$

holds for all categories $i = 1, \ldots, k$, due to the fact that the E_i appear in the denominator of the test statistic in Eq. (11.9).

Test decision: The rejection region for H_0 at significance level α is given by (right-sided test)

$$t_n > \chi^2_{k-1-r;1-\alpha} \, . \tag{11.11}$$

By Eq. (10.13), the p–value associated with a realisation t_n of (11.9) amounts to

$$p = P(T_n > t_n | H_0) = 1 - P(T_n \leq t_n | H_0) = 1 - \chi^2\texttt{cdf}(0, t_n, k-1-r) \, . \tag{11.12}$$

SPSS: Analyze → Nonparametric Tests → Legacy Dialogs → Chi-square ...

Note that in the spirit of **critical rationalism** the one-sample χ^2–goodness–of–fit–test provides a tool for empirically *excluding* possibilities of distribution laws for X.

11.3 One-sample t– and Z–tests for a population mean

The idea here is to test whether the unknown population mean μ of some continuous random variable X is equal to, less than, or greater than some reference value μ_0, to a given significance level α. To this end, it is required that X satisfy in the population $\mathbf{\Omega}$ a **Gaußian normal distribution**, i.e., $X \sim N(\mu; \sigma^2)$; cf. Sec. 8.5. The quantitative–analytical tool to be employed in this case is the parametric **one-sample t–test for a population mean** developed by Student [Gosset] (1908) [52], or, when the sample size $n \geq 50$, in consequence of the **central limit theorem** discussed in Sec. 8.13, the corresponding **one-sample Z–test**.

For an independent random sample $\boldsymbol{S_\Omega}$: $(X_1, \ldots, X_n)$ of size n, **normality** of the X-distribution can be tested for by a procedure due to the Russian mathematicians Andrey Nikolaevich Kolmogorov (1903–1987) and Nikolai Vasilyevich Smirnov (1900–1966), which tests the null hypothesis H_0: "There is no difference between the distribution of the sample data and a normal distribution" against the alternative H_1: "There is a difference between the distribution of the sample data and a normal distribution"; cf. Kolmogorov (1933) [25] and Smirnov (1939) [48]. This procedure is referred to as the **Kolmogorov–Smirnov–test** (or, for short, the KS–test). The associated test statistic evaluates the strength of the deviation of the empirical cumulative distribution function [cf. Eq. (2.4)] of given random sample data with sample mean $\bar{x}_n$ and sample variance s_n^2 from the `cdf` of a reference Gaußian normal distribution with parameters μ and σ^2 equal to these sample values [cf. Eq. (8.40)].

SPSS: Analyze → Nonparametric Tests → Legacy Dialogs → 1-Sample K-S . . . : Normal
R: `ks.test(`*variable*`, "pnorm")`

Formulated in a non-directed or a directed fashion, the starting point of the t–test resp. Z–test procedures are the

Hypotheses:

$$\begin{cases} H_0 : \mu = \mu_0 \quad \text{or} \quad \mu \geq \mu_0 \quad \text{or} \quad \mu \leq \mu_0 \\ H_1 : \mu \neq \mu_0 \quad \text{or} \quad \mu < \mu_0 \quad \text{or} \quad \mu > \mu_0 \end{cases} . \tag{11.13}$$

To measure the deviation of the sample data from the state conjectured to hold in the null hypothesis H_0, the difference between the sample mean $\bar{X}_n$ and the hypothesised population mean μ_0, normalised in analogy to Eq. (7.32) by the **standard error**

$$\frac{S_n}{\sqrt{n}} \tag{11.14}$$

of $\bar{X}_n$ given in Eq. (10.8), serves as the μ_0–dependent

Test statistic:

$$\boxed{T_n(X_1, \ldots, X_n) = \frac{\bar{X}_n - \mu_0}{S_n/\sqrt{n}} \overset{H_0}{\sim} \begin{cases} t(n-1) & \text{for} \quad n < 50 \\ \\ N(0;1) & \text{for} \quad n \geq 50 \end{cases}} , \tag{11.15}$$

which, under H_0, satisfies a t–distribution with $df = n-1$ degrees of freedom (cf. Sec. 8.7) resp. a standard normal distribution (cf. Sec. 8.5).

Test decision: Depending on the kind of test to be performed, the rejection region for H_0 at significance level α is given by

Kind of test	$\boldsymbol{H_0}$	$\boldsymbol{H_1}$	**Rejection region for $\boldsymbol{H_0}$**
(a) two-sided	$\mu = \mu_0$	$\mu \neq \mu_0$	$\lvert t_n\rvert > \begin{cases} t_{n-1;1-\alpha/2} & (t\text{–test}) \\ z_{1-\alpha/2} & (Z\text{–test}) \end{cases}$
(b) left-sided	$\mu \geq \mu_0$	$\mu < \mu_0$	$t_n < \begin{cases} t_{n-1;\alpha} = -t_{n-1;1-\alpha} & (t\text{–test}) \\ z_\alpha = -z_{1-\alpha} & (Z\text{–test}) \end{cases}$
(c) right-sided	$\mu \leq \mu_0$	$\mu > \mu_0$	$t_n > \begin{cases} t_{n-1;1-\alpha} & (t\text{–test}) \\ z_{1-\alpha} & (Z\text{–test}) \end{cases}$

p–values associated with realisations t_n of (11.15) can be obtained from Eqs. (10.11)–(10.13).

GDC: mode `STAT` → `TESTS` → `T-Test...` when $n < 50$, resp. mode `STAT` → `TESTS` → `Z-Test...` when $n \geq 50$.
SPSS: Analyze → Compare Means → One-Sample T Test ...
R: `t.test(`*variable*`, mu=`μ_0`)`,
`t.test(`*variable*`, mu=`μ_0`, alternative="less")`,
`t.test(`*variable*`, mu=`μ_0`, alternative="greater")`

Note: Regrettably, SPSS provides no option to select between a "one-tailed" (left-/right-sided) and a "two-tailed" (two-sided) t–test. The default setting is for a two-sided test. For the purpose of one-sided tests the p–value output of SPSS needs to be divided by 2.

11.4 One-sample χ^2–test for a population variance

In analogy to the statistical significance test described in the previous section 11.3, one may likewise test hypotheses on the value of an unknown population variance σ^2 with respect to a reference value σ_0^2 for a continuous random variable X which satisfies in $\boldsymbol{\Omega}$ a **Gaußian normal distribution**, i.e., $X \sim N(\mu; \sigma^2)$; cf. Sec. 8.5. These may also be formulated in a non-directed or directed fashion according to

Hypotheses:

$$\begin{cases} H_0 : \sigma^2 = \sigma_0^2 & \text{or} \quad \sigma^2 \geq \sigma_0^2 & \text{or} \quad \sigma^2 \leq \sigma_0^2 \\ H_1 : \sigma^2 \neq \sigma_0^2 & \text{or} \quad \sigma^2 < \sigma_0^2 & \text{or} \quad \sigma^2 > \sigma_0^2 \end{cases} . \tag{11.16}$$

In the **one-sample $\boldsymbol{\chi^2}$–test for a population variance**, the underlying σ_0^2–dependent

Test statistic:

$$\boxed{T_n(X_1,\ldots,X_n) = \frac{(n-1)S_n^2}{\sigma_0^2} \overset{H_0}{\sim} \chi^2(n-1)} \tag{11.17}$$

is chosen to be proportional to the sample variance defined by Eq. (10.7), and so, under H_0, satisfies a χ^2–distribution with $df = n-1$ degrees of freedom; cf. Sec. 8.6.

Test decision: Depending on the kind of test to be performed, the rejection region for H_0 at significance level α is given by

Kind of test	$\boldsymbol{H_0}$	$\boldsymbol{H_1}$	**Rejection region for $\boldsymbol{H_0}$**
(a) two-sided	$\sigma^2 = \sigma_0^2$	$\sigma^2 \neq \sigma_0^2$	$t_n \begin{cases} < \chi^2_{n-1;\alpha/2} \\ > \chi^2_{n-1;1-\alpha/2} \end{cases}$
(b) left-sided	$\sigma^2 \geq \sigma_0^2$	$\sigma^2 < \sigma_0^2$	$t_n < \chi^2_{n-1;\alpha}$
(c) right-sided	$\sigma^2 \leq \sigma_0^2$	$\sigma^2 > \sigma_0^2$	$t_n > \chi^2_{n-1;1-\alpha}$

p–values associated with realisations t_n of (11.17) can be obtained from Eqs. (10.11)–(10.13).

Regrettably, the one-sample χ^2–test for a population variance does not appear to have been implemented in the SPSS software package.

11.5 Two independent samples *t*–test for a population mean

Quantitative–empirical studies are frequently interested in the question as to what extent there exist significant differences between two subgroups of some population $\boldsymbol{\Omega}$ in the distribution of a metrically scaled variable X. Given that X is *normally distributed* in $\boldsymbol{\Omega}$ (cf. Sec. 8.5), the parametric **two independent samples *t*–test for a population mean** originating from work by Student [Gosset] (1908) [52] provides an efficient and powerful investigative tool.

For independent random samples of sizes $n_1, n_2 \geq 50$, normality of the X-distribution can be tested for by the **Kolmogorov–Smirnov–test**; cf. Sec. 11.3.

SPSS: Analyze → Nonparametric Tests → Legacy Dialogs → 1-Sample K-S . . . : Normal
R: `ks.test(`*variable*`, "pnorm")`

In addition, prior to the t–test procedure, one needs to establish whether or not the variances of X are significantly different in the two random samples selected. **Levene's test** provides an empirical method to test $H_0: \ \sigma_1^2 = \sigma_2^2$ against $H_1: \ \sigma_1^2 \neq \sigma_2^2$; cf. Levene (1960) [29].

R: `levene.test(`*variable~group variable*`)` (package: `car`)

The hypotheses of a t–test may be formulated in a non-directed fashion or in a directed one. Hence, the different kinds of conjectures are

Hypotheses: (test for differences)

$$\begin{cases} H_0 : \mu_1 - \mu_2 = 0 & \text{or} \quad \mu_1 - \mu_2 \geq 0 & \text{or} \quad \mu_1 - \mu_2 \leq 0 \\ H_1 : \mu_1 - \mu_2 \neq 0 & \text{or} \quad \mu_1 - \mu_2 < 0 & \text{or} \quad \mu_1 - \mu_2 > 0 \end{cases} . \tag{11.18}$$

A test statistic is constructed from the difference of sample means, $\bar{X}_{n_1} - \bar{X}_{n_2}$, standardised by the **standard error**

$$S_{\bar{X}_{n_1} - \bar{X}_{n_2}} := \sqrt{\frac{S_{n_1}^2}{n_1} + \frac{S_{n_2}^2}{n_2}} , \tag{11.19}$$

which derives from the associated theoretical **sampling distribution**. Thus, one obtains the

Test statistic:

$$\boxed{T_{n_1,n_2} := \frac{\bar{X}_{n_1} - \bar{X}_{n_2}}{S_{\bar{X}_{n_1} - \bar{X}_{n_2}}} \overset{H_0}{\sim} t(df) ,} \tag{11.20}$$

which, under H_0, is t–distributed (cf. Sec. 8.7) with a number of degrees of freedom determined by the relations

$$df := \begin{cases} n_1 + n_2 - 2 , & \text{when} \quad \sigma_1^2 = \sigma_2^2 \\ \dfrac{\left(\frac{S_{n_1}^2}{n_1} + \frac{S_{n_2}^2}{n_2}\right)^2}{\frac{(S_{n_1}^2/n_1)^2}{n_1 - 1} + \frac{(S_{n_2}^2/n_2)^2}{n_2 - 1}} , & \text{when} \quad \sigma_1^2 \neq \sigma_2^2 \end{cases} . \tag{11.21}$$

Test decision: Depending on the kind of test to be performed, the rejection region for H_0 at significance level α is given by

Kind of test	$\boldsymbol{H_0}$	$\boldsymbol{H_1}$	**Rejection region for** $\boldsymbol{H_0}$
(a) two-sided	$\mu_1 - \mu_2 = 0$	$\mu_1 - \mu_2 \neq 0$	$\lvert t_{n_1,n_2} \rvert > t_{df;1-\alpha/2}$
(b) left-sided	$\mu_1 - \mu_2 \geq 0$	$\mu_1 - \mu_2 < 0$	$t_{n_1,n_2} < t_{df;\alpha} = -t_{df;1-\alpha}$
(c) right-sided	$\mu_1 - \mu_2 \leq 0$	$\mu_1 - \mu_2 > 0$	$t_{n_1,n_2} > t_{df;1-\alpha}$

p–values associated with realisations t_{n_1,n_2} of (11.20) can be obtained from Eqs. (10.11)–(10.13).

GDC: mode `STAT` → `TESTS` → `2-SampTTest...`
SPSS: Analyze → Compare Means → Independent-Samples T Test ...

R: `t.test(`*variable~group variable*`)`,
`t.test(`*variable~group variable*`, alternative="less")`,
`t.test(`*variable~group variable*`, alternative="greater")`

Note: Regrettably, SPSS provides no option to select between a one-sided and a two-sided t–test. The default setting is for a two-sided test. For the purpose of one-sided tests the p–value output of SPSS needs to be divided by 2.

When the necessary conditions for the application of the independent sample t–test are not satisfied, the following alternative test procedures (typically of a weaker test power, though) for comparing two subgroups of $\mathbf{\Omega}$ with respect to the distribution of a metrically scaled variable X exist:

(i) at the **nominal** scale level, provided $E_{ij} \geq 5$ for all i, j, the **χ^2–test for homogeneity**; cf. Sec. 11.10 below, and

(ii) at the **ordinal** scale level, provided $n_1, n_2 \geq 8$, the two independent samples **Mann–Whitney–U–test** for a median; cf. the following Sec. 11.6.

11.6 Two independent samples Mann–Whitney–U–test for a population median

The non-parametric **two independent samples Mann–Whitney–U–test for a population median**, devised by the Austrian–US-American mathematician and statistician Henry Berthold Mann (1905–2000) and the US-American statistician Donald Ransom Whitney (1915–2001) in 1947 [36], can be applied to random sample data for ordinally scaled variables X, or for metrically scaled variables X which are *not* normally distributed in the population $\mathbf{\Omega}$. In both situations, the method employs **rank data** (cf. Sec. 4.3), which faithfully represents the original random sample data, to compare the medians of X between two independent groups. It aims to test empirically one of the following pairs of non-directed or directed

Hypotheses: (test for differences)

$$\begin{cases} H_0 : \tilde{x}_{0.5}(1) = \tilde{x}_{0.5}(2) & \text{or} \quad \tilde{x}_{0.5}(1) \geq \tilde{x}_{0.5}(2) \quad \text{or} \quad \tilde{x}_{0.5}(1) \leq \tilde{x}_{0.5}(2) \\ H_1 : \tilde{x}_{0.5}(1) \neq \tilde{x}_{0.5}(2) & \text{or} \quad \tilde{x}_{0.5}(1) < \tilde{x}_{0.5}(2) \quad \text{or} \quad \tilde{x}_{0.5}(1) > \tilde{x}_{0.5}(2) \end{cases} . \tag{11.22}$$

Given two independent sets of random sample data for X, **ranks** are being introduced on the basis of an ordered **joint random sample** of size $n = n_1 + n_2$ according to $x_i(1) \mapsto R[x_i(1)]$ and $x_i(2) \mapsto R[x_i(2)]$. From the ranks thus assigned to each of the two sets of data, one computes the

***U*–values:**

$$U_1 := n_1 n_2 + \frac{n_1(n_1+1)}{2} - \sum_{i=1}^{n_1} R[x_i(1)] \tag{11.23}$$

$$U_2 := n_1 n_2 + \frac{n_2(n_2+1)}{2} - \sum_{i=1}^{n_2} R[x_i(2)] \,, \tag{11.24}$$

which are subject to the identity $U_1 + U_2 = n_1 n_2$. Choose $U := \min(U_1, U_2)$.[1] For independent random samples of sizes $n_1, n_2 \geq 8$ (see, e.g., Bortz (2005) [4, p 151]), the standardised U–value serves as the

Test statistic:

$$\boxed{T_{n_1,n_2} := \frac{U - \mu_U}{\sigma_U} \overset{H_0}{\approx} N(0;1)\,,} \tag{11.25}$$

which, under H_0, approximately satisfies a standard normal distribution; cf. Sec. 8.5. Here, μ_U denotes the mean of the U–value expected under H_0; it is defined in terms of the sample sizes by

$$\mu_U := \frac{n_1 n_2}{2}\,; \tag{11.26}$$

while σ_U is the **standard error** of the U–value and can be obtained, e.g., from Bortz (2005) [4, Eq. (5.49)].

Test decision: Depending on the kind of test to be performed, the rejection region for H_0 at significance level α is given by

Kind of test	$\boldsymbol{H_0}$	$\boldsymbol{H_1}$	**Rejection region for $\boldsymbol{H_0}$**
(a) two-sided	$\tilde{x}_{0.5}(1) = \tilde{x}_{0.5}(2)$	$\tilde{x}_{0.5}(1) \neq \tilde{x}_{0.5}(2)$	$\lvert t_{n_1,n_2}\rvert > z_{1-\alpha/2}$
(b) left-sided	$\tilde{x}_{0.5}(1) \geq \tilde{x}_{0.5}(2)$	$\tilde{x}_{0.5}(1) < \tilde{x}_{0.5}(2)$	$t_{n_1,n_2} < z_\alpha = -z_{1-\alpha}$
(c) right-sided	$\tilde{x}_{0.5}(1) \leq \tilde{x}_{0.5}(2)$	$\tilde{x}_{0.5}(1) > \tilde{x}_{0.5}(2)$	$t_{n_1,n_2} > z_{1-\alpha}$

p–values associated with realisations t_{n_1,n_2} of (11.25) can be obtained from Eqs. (10.11)–(10.13).

SPSS: Analyze → Nonparametric Tests → Legacy Dialogs → 2 Independent Samples …: Mann-Whitney U
R: `wilcox.test(`*variable~group variable*`)`,
`wilcox.test(`*variable~group variable*`, alternative="less")`,
`wilcox.test(`*variable~group variable*`, alternative="greater")`

Note: Regrettably, SPSS provides no option to select between a one-sided and a two-sided U–test. The default setting is for a two-sided test. For the purpose of one-sided tests the p–value output of SPSS needs to be divided by 2.

[1] Since the U–values are tied to each other by the identity $U_1 + U_2 = n_1 n_2$, it makes no difference to this method when one chooses $U := \max(U_1, U_2)$ instead.

11.7 Two independent samples F–test for a population variance

In analogy to the independent samples t–test for a population mean of Sec. 11.5, one may likewise investigate for a metrically scaled variable X which satisfies a Gaußian normal distribution in $\mathbf{\Omega}$ (cf. Sec. 8.5) whether there exists a significant difference in the value of the population variance between two independent random samples.[2] The parametric **two independent samples F–test for a population variance** evaluates the non-directed resp. directed pairs of

Hypotheses: (test for differences)

$$\begin{cases} H_0 : \sigma_1^2 = \sigma_2^2 & \text{or} \quad \sigma_1^2 \geq \sigma_2^2 & \text{or} \quad \sigma_1^2 \leq \sigma_2^2 \\ H_1 : \sigma_1^2 \neq \sigma_2^2 & \text{or} \quad \sigma_1^2 < \sigma_2^2 & \text{or} \quad \sigma_1^2 > \sigma_2^2 \end{cases} . \tag{11.27}$$

Dealing with independent random samples of sizes n_1 and n_2, the ratio of the corresponding sample variances serves as a

Test statistic:

$$\boxed{T_{n_1,n_2} := \frac{S_{n_1}^2}{S_{n_2}^2} \overset{H_0}{\sim} F(n_1 - 1, n_2 - 1)\,,} \tag{11.28}$$

which, under H_0, satisfies an F–distribution with $df_1 = n_1 - 1$ and $df_2 = n_2 - 1$ degrees of freedom; cf. Sec. 8.8.

Test decision: Depending on the kind of test to be performed, the rejection region for H_0 at significance level α is given by

Kind of test	$\boldsymbol{H_0}$	$\boldsymbol{H_1}$	**Rejection region for $\boldsymbol{H_0}$**
(a) two-sided	$\sigma_1^2 = \sigma_2^2$	$\sigma_1^2 \neq \sigma_2^2$	$t_{n_1,n_2} \begin{cases} < 1/f_{n_2-1,n_1-1;1-\alpha/2} \\ > f_{n_1-1,n_2-1;1-\alpha/2} \end{cases}$
(b) left-sided	$\sigma_1^2 \geq \sigma_2^2$	$\sigma_1^2 < \sigma_2^2$	$t_{n_1,n_2} < 1/f_{n_2-1,n_1-1;1-\alpha}$
(c) right-sided	$\sigma_1^2 \leq \sigma_2^2$	$\sigma_1^2 > \sigma_2^2$	$t_{n_1,n_2} > f_{n_1-1,n_2-1;1-\alpha}$

p–values associated with realisations t_{n_1,n_2} of (11.28) can be obtained from Eqs. (10.11)–(10.13).

GDC: mode `STAT` → `TESTS` → `2-SampFTest...`
R: `var.test(`*variable~group variable*`)`,

[2] Run the Kolmogorov–Smirnov–test to check for normality of the distribution of X in the two random samples.

`var.test(`*variable~group variable*`, alternative="less"),`
`var.test(`*variable~group variable*`, alternative="greater")`

Regrettably, the two-sample F–test for a population variance does not appear to have been implemented in the SPSS software package. Instead, to address quantitative issues of the kind raised here, one may resort to **Levene's test**; cf. Sec. 11.5.

11.8 Two dependent samples *t*–test for a population mean

Besides investigating for significant differences in the distribution of a single variable X in two or more independent subgroups of some population $\mathbf{\Omega}$, many research projects are interested in finding out (i) how the distributional properties of a variable X have changed within one and the same random sample of $\mathbf{\Omega}$ in an experimental before–after situation, or (ii) how the distribution of a variable X differs between two subgroups of $\mathbf{\Omega}$ the sample units of which co-exist in a natural pairwise one-to-one correspondence to one another.

When the variable X in question is metrically scaled and satisfies a Gaußian normal distribution in $\mathbf{\Omega}$, significant differences can be tested for by means of the parametric **two dependent samples *t*–test for a population mean**. Denoting by A and B either some before and after instants, or the partners in a set of natural pairs (A, B), define for X the metrically scaled **difference variable**

$$D := X(A) - X(B) \; . \tag{11.29}$$

An *important test prerequisite* demands that D itself be *normally distributed* in $\mathbf{\Omega}$; cf. Sec. 8.5. Whether this property holds true can be checked via the **Kolmogorov–Smirnov–test**; cf. Sec. 11.3.

With μ_D denoting the population mean of the difference variable D, the

Hypotheses: (test for differences)

$$\begin{cases} H_0 : \mu_D = 0 \quad \text{or} \quad \mu_D \geq 0 \quad \text{or} \quad \mu_D \leq 0 \\ H_1 : \mu_D \neq 0 \quad \text{or} \quad \mu_D < 0 \quad \text{or} \quad \mu_D > 0 \end{cases} \tag{11.30}$$

can be given in a non-directed or a directed formulation. From the sample mean $\bar{D}$ and its associated **standard error**,

$$\frac{S_D}{\sqrt{n}} \; , \tag{11.31}$$

one obtains by means of standardisation according to Eq. (7.32) the

Test statistic:

$$\boxed{T_n := \frac{\bar{D}}{S_D/\sqrt{n}} \overset{H_0}{\sim} t(n-1) \; ,} \tag{11.32}$$

which, under H_0, satisfies a t–distribution with $df = n - 1$ degrees of freedom; cf. Sec. 8.7.

Test decision: Depending on the kind of test to be performed, the rejection region for H_0 at significance level α is given by

Kind of test	H_0	H_1	**Rejection region for H_0**
(a) two-sided	$\mu_D = 0$	$\mu_D \neq 0$	$\lvert t_n \rvert > t_{n-1;1-\alpha/2}$
(b) left-sided	$\mu_D \geq 0$	$\mu_D < 0$	$t_n < t_{n-1;\alpha} = -t_{n-1;1-\alpha}$
(c) right-sided	$\mu_D \leq 0$	$\mu_D > 0$	$t_n > t_{n-1;1-\alpha}$

p–values associated with realisations t_n of (11.32) can be obtained from Eqs. (10.11)–(10.13).

SPSS: Analyze → Compare Means → Paired-Samples T Test ...
R: `t.test(`*variableA*`,` *variableB*`, paired="T")`,
`t.test(`*variableA*`,` *variableB*`, paired="T", alternative="less")`,
`t.test(`*variableA*`,` *variableB*`, paired="T", alternative="greater")`

Note: Regrettably, SPSS provides no option to select between a one-sided and a two-sided t–test. The default setting is for a two-sided test. For the purpose of one-sided tests the p–value output of SPSS needs to be divided by 2.

11.9 Two dependent samples Wilcoxon–test for a population median

When the test prerequisites of the dependent samples t–test cannot be met, i.e., a given metrically scaled variable X cannot be assumed to satisfy a Gaußian normal distribution in $\mathbf{\Omega}$, or X is an ordinally scaled variable in the first place, the non-parametric **signed ranks test** published by the US-American chemist and statistician Frank Wilcoxon (1892–1965) in 1945 [62] constitutes a quantitative–empirical tool for comparing the distributional properties of X between two dependent random samples drawn from $\mathbf{\Omega}$. Like Mann and Whitney's U–test discussed in Sec. 11.6, it is built around the idea of **rank data** representing the original random sample data; cf. Sec. 4.3. Defining again a variable

$$D := X(A) - X(B)\,, \tag{11.33}$$

with associated median $\tilde{x}_{0.5}(D)$, the non-directed or directed pairs of

Hypotheses: (test for differences)

$$\begin{cases} H_0 : \tilde{x}_{0.5}(D) = 0 \quad \text{or} \quad \tilde{x}_{0.5}(D) \geq 0 \quad \text{or} \quad \tilde{x}_{0.5}(D) \leq 0 \\ H_1 : \tilde{x}_{0.5}(D) \neq 0 \quad \text{or} \quad \tilde{x}_{0.5}(D) < 0 \quad \text{or} \quad \tilde{x}_{0.5}(D) > 0 \end{cases} \tag{11.34}$$

need to be subjected to a suitable significance test.

For realisations d_i ($i = 1, \ldots, n$) of D, introduce **ranks** according to $d_i \mapsto R[|d_i|]$ for the ordered **absolute values** $|d_i|$, while keeping a record of the **sign** of each d_i. Exclude from the data set all null differences $d_i = 0$, leading to a sample of reduced size $n \mapsto n_{\rm red}$. Then form the **sums of ranks** W^+ for the $d_i > 0$ and W^- for the $d_i < 0$, respectively, which are linked to one another by the identity $W^+ + W^- = n_{\rm red}(n_{\rm red} + 1)/2$. Choose W^+.[3] For reduced sample sizes $n_{\rm red} > 20$ (see, e.g., Rinne (2008) [45, p 552]), one employs the

Test statistic:

$$\boxed{T_{n_{\rm red}} := \frac{W^+ - \mu_{W^+}}{\sigma_{W^+}} \overset{H_0}{\approx} N(0;1)\,,} \tag{11.35}$$

which, under H_0, approximately satisfies a standard normal distribution; cf. Sec. 8.5. Here, the mean μ_{W^+} expected under H_0 is defined in terms of $n_{\rm red}$ by

$$\mu_{W^+} := \frac{n_{\rm red}(n_{\rm red}+1)}{4}\,, \tag{11.36}$$

while the **standard error** σ_{W^+} can be computed from, e.g., Bortz (2005) [4, Eq. (5.52)].

Test decision: Depending on the kind of test to be performed, the rejection region for H_0 at significance level α is given by

Kind of test	$\boldsymbol{H_0}$	$\boldsymbol{H_1}$	**Rejection region for** $\boldsymbol{H_0}$
(a) two-sided	$\tilde{x}_{0.5}(D) = 0$	$\tilde{x}_{0.5}(D) \neq 0$	$\lvert t_{n_{\rm red}}\rvert > z_{1-\alpha/2}$
(b) left-sided	$\tilde{x}_{0.5}(D) \geq 0$	$\tilde{x}_{0.5}(D) < 0$	$t_{n_{\rm red}} < z_\alpha = -z_{1-\alpha}$
(c) right-sided	$\tilde{x}_{0.5}(D) \leq 0$	$\tilde{x}_{0.5}(D) >$	$t_{n_{\rm red}} > z_{1-\alpha}$

p–values associated with realisations $t_{n_{\rm red}}$ of (11.35) can be obtained from Eqs. (10.11)–(10.13).

SPSS: Analyze → Nonparametric Tests → Legacy Dialogs → 2 Related Samples . . . : Wilcoxon
R: `wilcox.test(`*variableA*`,` *variableB*`, paired="T")`,
`wilcox.test(`*variableA*`,` *variableB*`, paired="T", alternative="less")`,
`wilcox.test(`*variableA*`,` *variableB*`, paired="T", alternative="greater")`

Note: Regrettably, SPSS provides no option to select between a one-sided and a two-sided Wilcoxon–test. The default setting is for a two-sided test. For the purpose of one-sided tests the p–value output of SPSS needs to be divided by 2.

[3] Due to the identity $W^+ + W^- = n_{\rm red}(n_{\rm red}+1)/2$, choosing instead W^- would make no qualitative difference to the subsequent test procedure.

11.10 χ^2–test for homogeneity

Due to its independence of scale levels, the non-parametric **χ^2–test for homogeneity** constitutes the most generally applicable statistical test for significant differences in the distributional properties of a particular variable X between $k \in \mathbb{N}$ different subgroups of some population $\mathbf{\Omega}$. By assumption, the variable X may take values in a total of $l \in \mathbb{N}$ different **categories** a_j ($j = 1, \ldots, l$). Begin by formulating the

Hypotheses: (test for differences)

$$\begin{cases} H_0 : X \text{ satisfies the same distribution in all } k \text{ subgroups of } \mathbf{\Omega} \\ H_1 : X \text{ satisfies different distributions in at least two subgroups of } \mathbf{\Omega} \end{cases} . \tag{11.37}$$

With O_{ij} denoting the **observed frequency** of category a_j in subgroup i ($i = 1, \ldots, k$), and E_{ij} the under H_0 **expected frequency** of category a_j in subgroup i, the sum of rescaled squared **residuals** $\frac{(O_{ij} - E_{ij})^2}{E_{ij}}$ provides a useful

Test statistic:

$$\boxed{T_n := \sum_{i=1}^{k} \sum_{j=1}^{l} \frac{(O_{ij} - E_{ij})^2}{E_{ij}} \overset{H_0}{\approx} \chi^2[(k-1) \times (l-1)] .} \tag{11.38}$$

Under H_0, this test statistics satisfies approximately a χ^2–distribution with $df = (k-1) \times (l-1)$ degrees of freedom; cf. Sec. 8.6. The E_{ij} are defined in terms of **marginal observed frequencies** O_{i+} (the sum of observed frequencies in row i; cf. Eq. (4.3)) and O_{+j} (the sum of observed frequencies in column j; cf. Eq. (4.4)) by

$$E_{ij} := \frac{O_{i+} O_{+j}}{n} . \tag{11.39}$$

Note the *important (!) test prerequisite* that the total sample size $n := n_1 + \ldots + n_k$ be such that

$$E_{ij} \overset{!}{\geq} 5 \tag{11.40}$$

applies for all categories a_j and subgroups i.

Test decision: The rejection region for H_0 at significance level α is given by (right-sided test)

$$t_n > \chi^2_{(k-1)\times(l-1);1-\alpha} . \tag{11.41}$$

By Eq. (10.13), the p–value associated with a realisation t_n of (11.38) amounts to

$$p = P(T_n > t_n | H_0) = 1 - P(T_n \leq t_n | H_0) = 1 - \chi^2\texttt{cdf}\,(0, t_n, (k-1) \times (l-1)) . \tag{11.42}$$

GDC: mode `STAT` → `TESTS` → χ^2`-Test...`
SPSS: Analyze → Descriptive Statistics → Crosstabs ... → Statistics ...: Chi-square
R: `chisq.test(`*group variable*`,` *variable*`)`

Typically the power of a χ^2–test for homogeneity is weaker than for the related two procedures of comparing independent subgroups of $\mathbf{\Omega}$ which will be discussed in the subsequent Secs. 11.11 and 11.12.

11.11 One-way analysis of variance (ANOVA)

This powerful quantitative–analytical tool has been developed in the context of investigations on biometrical genetics by the English statistician Sir Ronald Aylmer Fisher FRS (1890–1962) (see Fisher (1918) [14]), and later extended by the US-American statistician Henry Scheffé (1907–1977) (see Scheffé (1959) [46]). It is of a parametric nature and can be interpreted alternatively as a method for[4]

(i) investigating the influence of a qualitative variable Y with $k \geq 3$ categories a_i ($i = 1, \ldots, k$), generally referred to as a "factor", on a quantitative variable X, or

(ii) testing for differences of the mean of a quantitative variable X between $k \geq 3$ different subgroups of some population $\mathbf{\Omega}$.

A necessary condition for the application of the **one-way analysis of variance (ANOVA)** test procedure is that the quantitative variable X to be investigated needs to be (a) *normally distributed* (cf. Sec. 8.5) in the $k \geq 3$ subgroups of the population $\mathbf{\Omega}$ considered, with, in addition, (b) *equal variances*. Both of these conditions also have to hold for each of a set of k mutually stochastically independent random variables $X_1, \ldots, X_k$ representing k random samples drawn independently from the identified k subgroups of $\mathbf{\Omega}$, of sizes $n_1, \ldots, n_k \in \mathbb{N}$, respectively. In the following, the element X_{ij} of the underlying $(n \times 2)$ data matrix $\boldsymbol{X}$ represents the jth value of X in the random sample drawn from the ith subgroup of $\mathbf{\Omega}$, with $\bar{X}_i$ the corresponding **subgroup sample mean**. The k independent random samples can be understood to form a **total random sample** of size $n := n_1 + \ldots + n_k = \sum_{i=1}^{k} n_i$, with **total sample mean** $\bar{X}_n$; cf. Eq. (10.6).

The intention of the ANOVA procedure in the variant (ii) stated above is to empirically test the

Hypotheses: (test for differences)

$$\begin{cases} H_0 : \mu_1 = \ldots = \mu_k = \mu_0 \\ H_1 : \mu_i \neq \mu_0 \text{ at least for one } i = 1, \ldots, k \end{cases} . \tag{11.43}$$

The necessary test prerequisites can be checked by (a) the **Kolmogorov–Smirnov–test** for normality of the X-distribution in each of the k subgroups of $\mathbf{\Omega}$ (cf. Sec. 11.3), and likewise by (b) **Levene's test** for $H_0 : \sigma_1^2 = \ldots = \sigma_k^2 = \sigma_0^2$ against H_1: "$\sigma_i^2 \neq \sigma_0^2$ at least for one $i = 1, \ldots, k$" to test for equality of the variances in these k subgroups (cf. Sec. 11.5).

R: `levene.test(`*variable~group variable*`)` (package: `car`)

The starting point of the ANOVA procedure is a simple algebraic decomposition of the **random sample values** X_{ij} into three additive components according to

$$X_{ij} = \bar{X}_n + (\bar{X}_i - \bar{X}_n) + (X_{ij} - \bar{X}_i) . \tag{11.44}$$

This expresses the X_{ij} as the sum of the total sample mean, $\bar{X}_n$, the deviation of the subgroup sample means from the total sample mean, $(\bar{X}_i - \bar{X}_n)$, and the residual deviation of the sample

[4] Only experimental designs with fixed effects are considered here.

values from their respective subgroup sample means, $(X_{ij} - \bar{X}_i)$. The decomposition of the X_{ij} motivates a **linear stochastic model** for the population $\mathbf{\Omega}$ of the form[5]

$$\text{in } \mathbf{\Omega}: \quad X_{ij} = \mu_0 + \alpha_i + \varepsilon_{ij} \tag{11.45}$$

in order to quantify, via the α_i ($i = 1, \ldots, k$), the potential influence of the qualitative variable Y on the quantitative variable X. Here μ_0 is the **population mean** of X, it holds that $\sum_{i=1}^{k} n_i \alpha_i = 0$, and it is assumed for the **random errors** ε_{ij} that $\varepsilon_{ij} \overset{\text{i.i.d.}}{\sim} N(0; \sigma_0^2)$, i.e., that they are identically normally distributed and mutually stochastically independent.

Having established the decomposition (11.44), one next turns to consider the associated set of **sums of squared deviations** defined by

$$\text{BSS} := \sum_{i=1}^{k} \sum_{j=1}^{n_i} \left(\bar{X}_i - \bar{X}_n\right)^2 = \sum_{i=1}^{k} n_i \left(\bar{X}_i - \bar{X}_n\right)^2 \tag{11.46}$$

$$\text{RSS} := \sum_{i=1}^{k} \sum_{j=1}^{n_i} \left(X_{ij} - \bar{X}_i\right)^2 \tag{11.47}$$

$$\text{TSS} := \sum_{i=1}^{k} \sum_{j=1}^{n_i} \left(X_{ij} - \bar{X}_n\right)^2 , \tag{11.48}$$

where the summations are (i) over all n_i sample units within a subgroup, and (ii) over all of the k subgroups themselves. The sums are referred to as, resp., (a) the sum of squared deviations between the subgroup samples (BSS), (b) the residual sum of squared deviations within the subgroup samples (RSS), and (c) the total sum of squared deviations (TSS) of the individual X_{ij} from the total sample mean $\bar{X}_n$. It is a fairly elaborate though straightforward algebraic exercise to show that these three squared deviation terms relate to one another according to the strikingly elegant identity (cf. Bosch (1999) [6, p 220f])

$$\text{TSS} = \text{BSS} + \text{RSS} . \tag{11.49}$$

Now, from the sums of squared deviations (11.46)–(11.48), one defines, resp., the **total sample variance**,

$$S^2_{\text{total}} := \frac{1}{n-1} \sum_{i=1}^{k} \sum_{j=1}^{n_i} \left(X_{ij} - \bar{X}_n\right)^2 = \frac{\text{TSS}}{n-1} , \tag{11.50}$$

involving $df = n - 1$ degrees of freedom, the **sample variance between subgroups**,

$$S^2_{\text{between}} := \frac{1}{k-1} \sum_{i=1}^{k} n_i \left(\bar{X}_i - \bar{X}_n\right)^2 = \frac{\text{BSS}}{k-1} , \tag{11.51}$$

[5] Formulated in the context of this linear stochastic model, the null and research hypotheses are $H_0 : \alpha_1 = \ldots = \alpha_k = 0$ and H_1: at least one $\alpha_i \neq 0$, respectively.

with $df = k - 1$, and the **mean sample variance within subgroups**,

$$S^2_{\text{within}} := \frac{1}{n-k} \sum_{i=1}^{k} \sum_{j=1}^{n_i} \left(X_{ij} - \bar{X}_i \right)^2 = \frac{\text{RSS}}{n-k} \,, \tag{11.52}$$

for which $df = n - k$.

Employing the latter two subgroup-specific dispersion measures, the set of hypotheses (12.11) may be recast into the alternative form

Hypotheses: (test for differences)

$$\begin{cases} H_0 : S^2_{\text{between}} \leq S^2_{\text{within}} \\ H_1 : S^2_{\text{between}} > S^2_{\text{within}} \end{cases} \,. \tag{11.53}$$

Finally, as a test statistic for the ANOVA procedure one chooses the ratio of variances[6]

$$T_{n,k} := \frac{(\text{sample variance between subgroups})}{(\text{mean sample variance within subgroups})} = \frac{\text{BSS}/(k-1)}{\text{RSS}/(n-k)} \,,$$

expressing the size of the "sample variance between subgroups" in terms of multiples of the "mean sample variance within subgroups"; it thus constitutes a relative measure. A real effect of difference between subgroups is thus given when the non-negative numerator turns out to be significantly larger than the non-negative denominator. Mathematically, this statistical measure of deviations between the data and the null hypothesis is given by

Test statistic:[7]

$$\boxed{T_{n,k} := \frac{S^2_{\text{between}}}{S^2_{\text{within}}} \overset{H_0}{\sim} F(k-1, n-k) \,.} \tag{11.54}$$

Under H_0, it satisfies an F–distribution with $df_1 = k - 1$ and $df_2 = n - k$ degrees of freedom; cf. Sec. 8.8.

It is a well-established standard in practical applications of the one-way ANOVA procedure to present the results in the form of a

Summary table:

ANOVA variability	sum of squares	df	mean square	test statistic
between groups	BSS	$k-1$	S^2_{between}	$t_{n,k}$
within groups	RSS	$n-k$	S^2_{within}	
total	TSS	$n-1$		

.

[6] This ratio is sometimes given as $T_{n,k} := \dfrac{(\text{explained variance})}{(\text{unexplained variance})}$, in analogy to expression (12.10) below. Occasionally, one also considers the coefficient $\eta^2 := \dfrac{\text{BSS}}{\text{TSS}}$, which, however, does not account for the degrees of freedom involved. In this respect, the modified coefficient $\tilde{\eta}^2 := \dfrac{S^2_{\text{between}}}{S^2_{\text{total}}}$ would constitute a more sophisticated measure.

[7] Note the one-to-one correspondence to the test statistic (11.28) employed in the independent samples F–test for a population variance.

Test decision: The rejection region for H_0 at significance level α is given by (right-sided test)

$$t_{n,k} > f_{k-1,n-k;1-\alpha} \, . \tag{11.55}$$

With Eq. (10.13), the p–value associated with a specific realisation $t_{n,k}$ of (11.54) amounts to

$$p = P(T_{n,k} > t_{n,k}|H_0) = 1 - P(T_{n,k} \leq t_{n,k}|H_0) = 1 - F\texttt{cdf}(0, t_{n,k}, k-1, n-k) \, . \tag{11.56}$$

GDC: mode `STAT` → `TESTS` → `ANOVA(`
SPSS: Analyze → Compare Means → One-Way ANOVA ...
R: `anova(lm(`*variable~group variable*`))` (variances equal),
`oneway.test(`*variable~group variable*`)` (variances not equal)

When a one-way ANOVA yields a statistically significant result, so-called **post-hoc tests** need to be run subsequently in order to identify those subgroups i whose means μ_i differ most drastically from the reference value μ_0. The **Student–Newman–Keuls–test** (Newman (1939) [37] and Keuls (1952) [23]), e.g., successively subjects the pairs of subgroups with the largest differences in sample means to independent samples t–tests; cf. Sec. 11.5. Other useful post-hoc tests are those developed by **Holm–Bonferroni** (Holm (1979) [22]), **Tukey** (Tukey (1977) [59]), or by **Scheffé** (Scheffé (1959) [46]).

SPSS: Analyze → Compare Means → One-Way ANOVA ... → Post Hoc ...
R: `pairwise.t.test(`*variable*`, `*group variable*`, p.adj="bonferroni")`

11.12 Kruskal–Wallis–test for a population median

Finally, a feasible alternative to the one-way ANOVA when the conditions for its legitimate application cannot be met, or one is interested in the distributional properties of a specific ordinally scaled variable X, is given by the non-parametric significance test devised by the US-American mathematician and statistician William Henry Kruskal (1919–2005) and the US-American economist and statistician Wilson Allen Wallis (1912–1998) in 1952 [26]. The **Kruskal–Wallis–test** serves to detect significant differences for a population median of an ordinally or metrically scaled variable X between $k \geq 3$ independent subgroups of some population $\mathbf{\Omega}$. To be investigated is the pair of mutually exclusive

Hypotheses: (test for differences)

$$\begin{cases} H_0 : \tilde{x}_{0.5}(1) = \ldots = \tilde{x}_{0.5}(k) \\ H_1 : \text{at least one } \tilde{x}_{0.5}(i) \; (i = 1, \ldots, k) \text{ is different from the other group medians} \end{cases} \, . \tag{11.57}$$

Introduce **ranks** according to $x_j(1) \mapsto R[x_j(1)]$, ..., and $x_j(k) \mapsto R[x_j(k)]$ within the random samples drawn independently from each of the $k \geq 3$ subgroups of $\mathbf{\Omega}$ on the basis of an ordered **joint random sample** of size $n := n_1 + \ldots + n_k = \sum_{i=1}^{k} n_i$; cf. Sec. 4.3. Then form the **sum of ranks** for each random sample separately, i.e.,

$$R_{+i} := \sum_{j=1}^{n_i} R[x_j(i)] \qquad (i = 1, \ldots, k) \, . \tag{11.58}$$

Provided the sample sizes satisfy the condition $n_i \geq 5$ for all $k \geq 3$ independent random samples (hence, $n \geq 15$), the test procedure can be based on the

Test statistic:

$$\boxed{T_{n,k} := \left[\frac{12}{n(n+1)} \sum_{i=1}^{k} \frac{R_{+i}^2}{n_i} \right] - 3(n+1) \overset{H_0}{\approx} \chi^2(k-1) \,,} \tag{11.59}$$

which, under H_0, approximately satisfies a χ^2–distribution with $df = k-1$ degrees of freedom (cf. Sec. 8.6); see, e.g., Rinne (2008) [45, p 553].

Test decision: The rejection region for H_0 at significance level α is given by (right-sided test)

$$t_{n,k} > \chi^2_{k-1;1-\alpha} \,. \tag{11.60}$$

By Eq. (10.13), the p–value associated with a realisation $t_{n,k}$ of (11.59) amounts to

$$p = P(T_{n,k} > t_{n,k}|H_0) = 1 - P(T_{n,k} \leq t_{n,k}|H_0) = 1 - \chi^2\texttt{cdf}(0, t_{n,k}, k-1) \,. \tag{11.61}$$

SPSS: Analyze → Nonparametric Tests → Legacy Dialogs → K Independent Samples …: Kruskal-Wallis H
R: `kruskal.test(`*variable~group variable*`)`

Chapter 12

Bivariate methods of statistical data analysis: testing for association

Recognising patterns of regularity in the variability of data sets for given (observable) variables, and explaining them in terms of **causal relationships** in the context of a suitable **theoretical model**, is one of the main objectives of any empirical scientific discipline; see, e.g., Penrose (2004) [43]. Causal relationships are viewed as intimately related to **interactions** between objects or agents of the physical or/and of the social kind. A *necessary condition* on the way of theoretically fathoming causal relationships is to establish empirically the existence of significant **statistical associations** between the variables in question. The possibility of replication of observational or experimental results of this kind, when given, lends strong support in favour of this idea. Regrettably, however, the existence of causal relationships between two variables *cannot* be established with absolute certainty by empirical means; compelling theoretical arguments need to take over. Causal relationships between variables imply an unambiguous distinction between **independent variables** and **dependent variables**. In the following, we will discuss the principles of the simplest three inferential statistical methods providing empirical checks of the aforementioned necessary condition in the **bivariate case**.

12.1 Correlation analysis and simple linear regression

12.1.1 t–test for a correlation

The parametric **correlation analysis** presupposes metrically scaled variables X and Y which satisfy a **bivariate normal distribution** in Ω. Its aim is to investigate whether or not X and Y feature a quantitative–statistical association of a *linear* nature, given random sample data $\boldsymbol{X} \in \mathbb{R}^{n\times 2}$. Formulated in terms of the **population correlation coefficient** ρ according to Auguste Bravais (1811–1863) and Karl Pearson FRS (1857–1936), the method tests H_0 against H_1 in one of the alternative pairs of

Hypotheses: (test for association)

$$\begin{cases} H_0 : \rho = 0 & \text{or} \quad \rho \geq 0 & \text{or} \quad \rho \leq 0 \\ H_1 : \rho \neq 0 & \text{or} \quad \rho < 0 & \text{or} \quad \rho > 0 \end{cases} , \tag{12.1}$$

with $-1 \leq \rho \leq +1$.

Normality of the marginal X- and Y-distributions in a given random sample $\boldsymbol{S_\Omega}$: $(X_1, \ldots, X_n;$-$Y_1, \ldots, Y_n)$ drawn from $\boldsymbol{\Omega}$ can again be tested for (for $n \geq 50$) by the **Kolmogorov–Smirnov–test**; cf. Sec. 11.3. A **scatter plot** of the raw sample data $\{(x_i, y_i)\}_{i=1,\ldots,n}$ represents features of the **joint $(\boldsymbol{X}, \boldsymbol{Y})$-distribution**.

SPSS: Analyze → Nonparametric Tests → Legacy Dialogs → 1-Sample K-S ...: Normal
R: `ks.test(`*variable*`, "pnorm")`

Rescaling the **sample correlation coefficient** r of Eq. (4.19) by the inverse of the **standard error** of r,

$$\sqrt{\frac{1-r^2}{n-2}}\,, \tag{12.2}$$

which can be obtained from the theoretical **sampling distribution** of r, presently yields the

Test statistic:

$$\boxed{T_n := \sqrt{n-2}\,\frac{r}{\sqrt{1-r^2}} \overset{H_0}{\sim} t(n-2)\,,} \tag{12.3}$$

which, under H_0, satisfies a t–distribution with $df = n-2$ degrees of freedom; cf. Sec. 8.7.

Test decision: Depending on the kind of test to be performed, the rejection region for H_0 at significance level α is given by

Kind of test	$\boldsymbol{H_0}$	$\boldsymbol{H_1}$	**Rejection region for $\boldsymbol{H_0}$**
(a) two-sided	$\rho = 0$	$\rho \neq 0$	$\lvert t_n \rvert > t_{n-2;1-\alpha/2}$
(b) left-sided	$\rho \geq 0$	$\rho < 0$	$t_n < t_{n-2;\alpha} = -t_{n-2;1-\alpha}$
(c) right-sided	$\rho \leq 0$	$\rho > 0$	$t_n > t_{n-2;1-\alpha}$

p–values associated with realisations t_n of (12.3) can be obtained from Eqs. (10.11)–(10.13).

SPSS: Analyze → Correlate → Bivariate ...: Pearson
R: `cor.test(`*variable1*`, `*variable2*`)`,
`cor.test(`*variable1*`, `*variable2*`, alternative="less")`,
`cor.test(`*variable1*`, `*variable2*`, alternative="greater")`

It is generally recommended to handle significant test results of a **correlation analysis** for metrically scaled variables X and Y with some care due to the possibility of **spurious correlations** induced by additional **control variables** $Z, \ldots$ acting in the background. To exclude this possibility, the correlation analysis should be repeated for homogeneous subgroups of the sample $\boldsymbol{S_\Omega}$.

12.1.2 F–test of a regression model

For correlations between metrically scaled variables X and Y significant in $\mathbf{\Omega}$ at level α, where ρ takes a magnitude in the interval

$$0.6 \leq |\rho| \leq 1.0 \,,$$

it is meaningful to ask which *linear* mathematical model best represents the detected linear statistical association; cf. Pearson (1903) [42]. To this end, **simple linear regression** seeks to devise a **linear stochastic regression model** for the population $\mathbf{\Omega}$ of the form

$$\text{in } \mathbf{\Omega}: \quad Y_i = \alpha + \beta x_i + \varepsilon_i \qquad (i = 1, \ldots, n) \,, \tag{12.4}$$

which, for instance, assigns X the role of an **independent variable** (and so its values x_i can be considered prescribed by the modeller) and Y the role of a **dependent variable**, thus rendering the model essentially univariate in nature. The **regression coefficients** α and β denote the unknown $\boldsymbol{y}$**–intercept** and **slope** of the model in $\mathbf{\Omega}$. For the **random errors** ε_i it is assumed that

$$\varepsilon_i \overset{\text{i.i.d.}}{\sim} N(0; \sigma^2) \,, \tag{12.5}$$

meaning they are identically normally distributed (with mean zero and constant variance σ^2) and mutually stochastically independent. With respect to the random sample $\boldsymbol{S}_{\mathbf{\Omega}}$: $(X_1, \ldots, X_n;$-$Y_1, \ldots, Y_n)$, the supposed linear relationship between X and Y is expressed by

$$\text{in } \boldsymbol{S}_{\mathbf{\Omega}}: \quad y_i = a + b x_i + e_i \qquad (i = 1, \ldots, n) \,. \tag{12.6}$$

Residuals are thereby defined according to

$$e_i := y_i - \hat{y}_i = y_i - a - b x_i \qquad (i = 1, \ldots, n) \,, \tag{12.7}$$

which, for given value of x_i, encode the difference between the observed realisations y_i and the corresponding (by the linear regression model) predicted realisations $\hat{y}_i$ of Y. By construction the residuals satisfy the condition $\displaystyle\sum_{i=1}^{n} e_i = 0$.

Next, introduce **sums of squared deviations** for the Y-data in line with the ANOVA procedure of Sec. 11.11, i.e.,

$$\text{TSS} \;:=\; \sum_{i=1}^{n} (y_i - \bar{y})^2 \tag{12.8}$$

$$\text{RSS} \;:=\; \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \;=\; \sum_{i=1}^{n} e_i^2 \,. \tag{12.9}$$

In terms of these quantities, the **coefficient of determination** of Eq. (5.8) for assessing the **goodness-of-the-fit** of a regression model can be expressed by

$$B = \frac{\text{TSS} - \text{RSS}}{\text{TSS}} = \frac{(\text{total variance of } Y) - (\text{unexplained variance of } Y)}{(\text{total variance of } Y)} \,, \tag{12.10}$$

where the latter equality holds for simple linear regression (with just a single independent variable) only. The normalised measure B, of range $0 \leq B \leq 1$, expresses the proportion of variability in a data set of Y which can be explained by the corresponding variability of X through the **best-fit regression model**.

In the methodology of a **regression analysis**, the first issue to be addressed is to test the significance of the overall **regression model** (12.4), i.e., to test H_0 against H_1 in the set of

Hypotheses: (test for differences)

$$\begin{cases} H_0 : \beta = 0 \\ H_1 : \beta \neq 0 \end{cases} . \tag{12.11}$$

Exploiting the goodness-of-the-fit aspect of the regression model quantified by the coefficient of determination (12.10), one derives the (see, e.g., Hatzinger and Nagel (2009) [20, Eq. (7.8)])

Test statistic:[1]

$$\boxed{T_n := (n-2)\,\frac{B}{1-B} \overset{H_0}{\sim} F(1, n-2)\,,} \tag{12.12}$$

which, under H_0, satisfies an F–distribution with $df_1 = 1$ and $df_2 = n - 2$ degrees of freedom; cf. Sec. 8.8.

Test decision: The rejection region for H_0 at significance level α is given by (right-sided test)

$$t_n > f_{1,n-2;1-\alpha}\,. \tag{12.13}$$

With Eq. (10.13), the p–value associated with a specific realisation t_n of (12.12) amounts to

$$p = P(T_n > t_n|H_0) = 1 - P(T_n \leq t_n|H_0) = 1 - F\texttt{cdf}(0, t_n, 1, n-2)\,. \tag{12.14}$$

12.1.3 *t*–test for the regression coefficients

The second issue to be addressed in a systematic **regression analysis** is to test statistically which of the regression coefficients in Eq. (12.4) is *non-zero*. In the case of simple linear regression, though, the matter for the coefficient β is settled already by the ***F*–test** of the regression model just outlined, resp. the ***t*–test** for ρ described in Subsec. 12.1.1; see, e.g., Levin *et al* (2010) [30, p 389f]. However, when extending the regression analysis of data to the more complex case of **multiple linear regression**, an approach frequently employed in the research literature of the **Social Sciences** and **Economics**, this question attains relevance in its own right. In view of this prospect, we continue with our methodological considerations.

First of all, **unbiased point estimators** for the regression coefficients α and β in Eq. (12.4) are obtained from application to the data of Gauß' method of **minimising the sum of squared residuals** (RSS) (cf. Gauß (1809) [17] and Ch. 5),

$$\text{minimise}\left(\text{RSS} = \sum_{i=1}^{n} e_i^2\right)\,,$$

[1] Note that with the identity $B = r^2$ of Eq. (5.9), which applies in simple linear regression, this is just the square of the test statistic (12.3).

yielding

$$b = \frac{S_Y}{s_X}\, r \qquad \text{and} \qquad a = \bar{Y} - b\bar{x}\ . \tag{12.15}$$

The equation of the **best-fit linear regression model** is thus given by

$$\boxed{\hat{y} = \bar{Y} + \frac{S_Y}{s_X}\, r\,(x - \bar{x})\ ,} \tag{12.16}$$

and can be employed for purposes of generating predictions for Y, given an independent value of X in the empirical interval $[x_{(1)}, x_{(n)}]$.

Next, the **standard errors** associated with the values of the point estimators a and b in Eq. (12.15) are derived from the corresponding theoretical **sampling distributions** and amount to (cf., e.g., Hartung *et al* (2005) [21, p 576ff])

$$\text{SE}_a \ := \ \sqrt{\frac{1}{n} + \frac{\bar{x}}{(n-1)s_X^2}}\,\text{SE}_e \tag{12.17}$$

$$\text{SE}_b \ := \ \frac{\text{SE}_e}{\sqrt{n-1}\,s_X}\ , \tag{12.18}$$

where the **standard error of the residuals** e_i is defined by

$$\text{SE}_e := \sqrt{\frac{\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2}{n-2}}\ . \tag{12.19}$$

We now describe the test procedure for the **regression coefficient** β. To be tested is H_0 against H_1 in one of the alternative pairs of

Hypotheses: (test for differences)

$$\begin{cases} H_0 : \beta = 0 & \text{or} \quad \beta \geq 0 & \text{or} \quad \beta \leq 0 \\ H_1 : \beta \neq 0 & \text{or} \quad \beta < 0 & \text{or} \quad \beta > 0 \end{cases}\ . \tag{12.20}$$

Dividing the **sample regression slope** b by its **standard error** yields the

Test statistic:

$$\boxed{T_n := \frac{b}{\text{SE}_b} \ \overset{H_0}{\sim}\ t(n-2)\ ,} \tag{12.21}$$

which, under H_0, satisfies a t–distribution with $df = n - 2$ degrees of freedom; cf. Sec. 8.7.

Test decision: Depending on the kind of test to be performed, the rejection region for H_0 at significance level α is given by

Kind of test	$\boldsymbol{H_0}$	$\boldsymbol{H_1}$	**Rejection region for** $\boldsymbol{H_0}$
(a) two-sided	$\beta = 0$	$\beta \neq 0$	$\lvert t_n \rvert > t_{n-2;1-\alpha/2}$
(b) left-sided	$\beta \geq 0$	$\beta < 0$	$t_n < t_{n-2;\alpha} = -t_{n-2;1-\alpha}$
(c) right-sided	$\beta \leq 0$	$\beta > 0$	$t_n > t_{n-2;1-\alpha}$

p–values associated with realisations t_n of (12.21) can be obtained from Eqs. (10.11)–(10.13). We emphasise once more that for simple linear regression the test procedure just described is equivalent to the **correlation analysis** of Subsec. 12.1.1.

An analogous ***t*–test** needs to be run to check whether the **regression coefficient** α is non-zero, too, using the ratio $\dfrac{a}{\mathrm{SE}_a}$ as the test statistic. However, in particular when the origin of X is not contained in the empirical interval $[x_{(1)}, x_{(n)}]$, the null hypothesis $H_0 : \alpha = 0$ is a meaningless statement.

GDC: mode `STAT` → `TESTS` → `LinRegTTest...`
SPSS: Analyze → Regression → Linear .....
R: `summary(lm(`*variable1~variable2*`))`

Note: Regrettably, SPSS provides no option to select between a one-sided and a two-sided t–test. The default setting is for a two-sided test. For the purpose of one-sided tests the p–value output of SPSS needs to be divided by 2.

Lastly, by means of an **analysis of the residuals** one can assess the extent to which the prerequisites of a regression analysis stated in Eq. (12.5) are satisfied:

(i) for $n \geq 50$, **normality** of the distribution of **residuals** e_i ($i = 1, \ldots, n$) can be checked by means of a **Kolmogorov–Smirnov–test**; cf. Sec. 11.3;

(ii) **homoscedasticity** of the e_i ($i = 1, \ldots, n$), i.e., whether or not they have constant variance, can be investigated qualitatively in terms of a **scatter plot** that marks the standardised e_i (along the vertical axis) against the corresponding predicted Y-values $\hat{y}_i$ ($i = 1, \ldots, n$) (along the horizontal axis). A circularly shaped envelope of the cloud of points indicates that homoscedasticity applies.

In reality, many quantitative phenomena studied in the **Natural Sciences** and in the **Social Sciences** prove to be of an inherently **non-linear nature**; see e.g. Gleick (1987) [19], Penrose (2004) [43], and Smith (2007) [49]. On the one hand, this increases the level of complexity involved in the data analysis, on the other, non-linear processes offer the reward of a plethora of interesting and intriguing (dynamical) phenomena.

12.2 Rank correlation analysis

When the variables X and Y are metrically scaled but *not* normally distributed in the population $\mathfrak{Omega}$, or X and Y are ordinally scaled in the first place, the standard tool for testing for a statistical association between X and Y is the parametric **rank correlation analysis** developed by the English psychologist and statistician Charles Edward Spearman FRS (1863–1945) in 1904 [51]. This approach, like the univariate test procedures of Mann and Whitney, Wilcoxon, and Kruskal and Wallis discussed in Ch. 11, is again fundamentally rooted in the concept of **ranks** representing statistical data which have a natural order, introduced in Sec. 4.3.

Following the translation of the original data pairs into corresponding **rank data pairs**,

$$(x_i, y_i) \mapsto [R(x_i), R(y_i)] \qquad (i = 1, \ldots, n) \;, \tag{12.22}$$

the objective is to subject H_0 in the alternative sets of

Hypotheses: (test for association)

$$\begin{cases} H_0 : \rho_S = 0 \quad \text{or} \quad \rho_S \geq 0 \quad \text{or} \quad \rho_S \leq 0 \\ H_1 : \rho_S \neq 0 \quad \text{or} \quad \rho_S < 0 \quad \text{or} \quad \rho_S > 0 \end{cases} \;, \tag{12.23}$$

with ρ_S ($-1 \leq \rho_S \leq +1$) the **population rank correlation coefficient**, to a test of statistical significance at level α. Provided the size of the random sample is such that $n \geq 30$ (see, e.g., Bortz (2005) [4, p 233]), there exists a suitable

Test statistic:

$$\boxed{T_n := \sqrt{n-2}\,\frac{r_S}{\sqrt{1-r_S^2}} \overset{H_0}{\approx} t(n-2)\;,} \tag{12.24}$$

which, under H_0, approximately satisfies a t–distribution with $df = n - 2$ degrees of freedom; cf. Sec. 8.7. Here, r_S denotes the **sample rank correlation coeffcient** defined in Eq. (4.31).

Test decision: Depending on the kind of test to be performed, the rejection region for H_0 at significance level α is given by

Kind of test	$\boldsymbol{H_0}$	$\boldsymbol{H_1}$	**Rejection region for** $\boldsymbol{H_0}$
(a) two-sided	$\rho_S = 0$	$\rho_S \neq 0$	$\lvert t_n \rvert > t_{n-2;1-\alpha/2}$
(b) left-sided	$\rho_S \geq 0$	$\rho_S < 0$	$t_n < t_{n-2;\alpha} = -t_{n-2;1-\alpha}$
(c) right-sided	$\rho_S \leq 0$	$\rho_S > 0$	$t_n > t_{n-2;1-\alpha}$

p–values associated with realisations t_n of (12.24) can be obtained from Eqs. (10.11)–(10.13).

SPSS: Analyze → Correlate → Bivariate ...: Spearman
R: `cor.test(`*variable1*`, `*variable2*`, method="spearman"),`
`cor.test(`*variable1*`, `*variable2*`, method="spearman", alternative="less"),`
`cor.test(`*variable1*`, `*variable2*`, method="spearman",`
`alternative="greater")`

12.3 χ^2–test for independence

The non-parametric **χ^2–test for independence** constitutes the most generally applicable significance test for bivariate statistical associations. Due to its formal indifference to the scale levels of the variables X and Y involved in the investigation, it may be used for statistical analysis of any kind of pairwise combinations between nominally, ordinally and metrically scaled variables. The advantage of generality of the method is paid for at the price of a typically weaker test power.

Given qualitative and/or quantitative variables X and Y that take values in a spectrum of k mutually exclusive categories $a_1, \ldots, a_k$ resp. l categories $b_1, \ldots, b_l$, the intention is to subject H_0 in the pair of alternative

Hypotheses: (test for association)

$$\begin{cases} H_0 : \text{There does not exist a statistical association between } X \text{ and } Y \text{ in } \mathbf{\Omega} \\ H_1 : \text{There does exist a statistical association between } X \text{ and } Y \text{ in } \mathbf{\Omega} \end{cases} \tag{12.25}$$

to a convenient empirical significance test at level α.

A conceptual issue that requires special attention along the way is the definition of a reasonable "zero point" on the scale of statistical dependence of variables X and Y (which one aims to establish). This problem is solved by recognising that a common feature of sample data for variables of all scale levels is the information residing in the distribution of (relative) frequencies over (all possible combinations of) categories and drawing an analogy to the concept of stochastic independence of two events as defined in probability theory by Eq. (6.13). In this way, by definition we refer to variables X and Y as being mutually **statistically independent** provided that the relative frequencies h_{ij} of *all* bivariate combinations of categories (a_i, b_j) are numerically equal to the products of the univariate marginal relative frequencies h_{i+} of a_i and h_{+j} of b_j (cf. Sec. 4.1), i.e.,

$$h_{ij} = h_{i+} h_{+j} \; . \tag{12.26}$$

Translated into the language of random sample variables, viz. introducing **sample observed frequencies**, this operational **independence condition** is re-expressed by $O_{ij} = E_{ij}$, where the O_{ij} denote the **observed frequencies** of all bivariate category combinations (a_i, b_j) in a cross tabulation underlying a specific random sample of size n, and the quantities E_{ij}, which are defined in terms of (i) the univariate sum O_{i+} of observed frequencies in row i, see Eq. (4.3), (ii) the univariate sum O_{+j} of observed frequencies in column j, see Eq. (4.4), and (iii) the sample size n by $E_{ij} := \dfrac{O_{i+} O_{+j}}{n}$, are interpreted as the **expected frequencies** of (a_i, b_j) given that X and Y

are statistically independent. Expressing deviations between observed and (under independence) expected frequencies via the **residuals** $O_{ij} - E_{ij}$, the hypotheses may be reformulated as

Hypotheses: (test for association)

$$\begin{cases} H_0 : O_{ij} - E_{ij} = 0 \\ H_1 : O_{ij} - E_{ij} \neq 0 \end{cases} . \tag{12.27}$$

For the subsequent test procedure to be reliable, it is *very important (!)* that the empirical prerequisite

$$E_{ij} \overset{!}{\geq} 5 \tag{12.28}$$

holds for all values of $i = 1 \ldots, k$ and $j = 1, \ldots, l$ such that one avoids the possibility that individual rescaled squared residuals $\dfrac{(O_{ij} - E_{ij})^2}{E_{ij}}$ become artificially magnified. The latter constitute the core of the

Test statistic:

$$\boxed{T_n := \sum_{i=1}^{k} \sum_{j=1}^{l} \frac{(O_{ij} - E_{ij})^2}{E_{ij}} \overset{H_0}{\approx} \chi^2[(k-1) \times (l-1)] \,,} \tag{12.29}$$

which, under H_0, approximately satisfies a χ^2–distribution with $df = (k-1) \times (l-1)$ degrees of freedom; cf. Sec. 8.6.

Test decision: The rejection region for H_0 at significance level α is given by (right-sided test)

$$t_n > \chi^2_{(k-1)\times(l-1);1-\alpha} \,. \tag{12.30}$$

By Eq. (10.13), the p–value associated with a realisation t_n of (12.29) amounts to

$$p = P(T_n > t_n | H_0) = 1 - P(T_n \leq t_n | H_0) = 1 - \chi^2\texttt{cdf}\,(0, t_n, (k-1) \times (l-1)) \,. \tag{12.31}$$

GDC: mode `STAT` → `TESTS` → χ^2`-Test...`
SPSS: Analyze → Descriptive Statistics → Crosstabs ... → Statistics ...: Chi-square
R: `chisq.test(`*row variable*`,` *column variable*`)`

The χ^2–test for independence can establish the **existence** of a significant association between variables X and Y. The **strength** of the association, on the other hand, may be measured in terms of **Cramér's *V*** (Cramér (1946) [8]), which has a normalised range of values given by $0 \leq V \leq 1$; cf. Eq. (4.35) and Sec. 4.4. Low values of V in the case of significant associations between variables X and Y typically indicate the statistical influence of additional **control variables**.

SPSS: Analyze → Descriptive Statistics → Crosstabs ... → Statistics ...: Phi and Cramer's V

Appendix A

Principle component analysis of a (2×2) correlation matrix

Consider a real-valued $(\mathbf{2 \times 2})$ **correlation matrix** expressed by

$$\boldsymbol{R} = \begin{pmatrix} 1 & r \\ r & 1 \end{pmatrix} , \qquad -1 \leq r \leq +1 , \tag{A.1}$$

which, by construction, is symmetric. Its **trace** amounts to $\mathrm{Tr}(\boldsymbol{R}) = 2$, while its **determinant** is $\det(\boldsymbol{R}) = 1 - r^2$. Consequently, $\boldsymbol{R}$ is regular as long as $r \neq \pm 1$. We seek to determine the **eigenvalues** (or **principle components**) and corresponding **eigenvectors** (or directions of the **principle axes**) of $\boldsymbol{R}$, i.e., real numbers λ and real-valued vectors $\boldsymbol{v}$ such that the condition

$$\boldsymbol{R}\,\boldsymbol{v} \stackrel{!}{=} \lambda\,\boldsymbol{v} \qquad \Leftrightarrow \qquad (\boldsymbol{R} - \lambda \mathbf{1})\,\boldsymbol{v} \stackrel{!}{=} \mathbf{0} \tag{A.2}$$

applies. Solution of this algebraic problem leads to the **characteristic equation**

$$0 \stackrel{!}{=} \det(\boldsymbol{R} - \lambda \mathbf{1}) = (1-\lambda)^2 - r^2 = (\lambda - 1)^2 - r^2 . \tag{A.3}$$

Hence, it is clear that $\boldsymbol{R}$ possesses the two **eigenvalues**

$$\lambda_1 = 1 + r \qquad \text{and} \qquad \lambda_2 = 1 - r , \tag{A.4}$$

showing that $\boldsymbol{R}$ is **positive-definite** whenever $|r| < 1$. The normalised **eigenvectors** associated with λ_1 and λ_2, obtained from Eq. (A.2), are then

$$\boldsymbol{v}_1 = \frac{1}{\sqrt{2}} \begin{pmatrix} 1 \\ 1 \end{pmatrix} \qquad \text{and} \qquad \boldsymbol{v}_2 = \frac{1}{\sqrt{2}} \begin{pmatrix} -1 \\ 1 \end{pmatrix} , \tag{A.5}$$

and constitute a right-handedly oriented basis of the two-dimensional **eigenspace** of $\boldsymbol{R}$. Note that due to the symmetry of $\boldsymbol{R}$ it holds that $\boldsymbol{v}_1^T \cdot \boldsymbol{v}_2 = 0$.

The normalised eigenvectors of $\boldsymbol{R}$ define a regular orthogonal **transformation matrix** $\boldsymbol{M}$, with inverse $\boldsymbol{M}^{-1} = \boldsymbol{M}^T$, given by

$$\boldsymbol{M} = \frac{1}{\sqrt{2}} \begin{pmatrix} 1 & -1 \\ 1 & 1 \end{pmatrix} \qquad \text{and} \qquad \boldsymbol{M}^{-1} = \frac{1}{\sqrt{2}} \begin{pmatrix} 1 & 1 \\ -1 & 1 \end{pmatrix} = \boldsymbol{M}^T , \tag{A.6}$$

where $\mathrm{Tr}(\boldsymbol{M}) = \sqrt{2}$ and $\det(\boldsymbol{M}) = 1$. The correlation matrix $\boldsymbol{R}$ can now be **diagonalised** by means of a rotation with $\boldsymbol{M}$ according to[1]

$$\begin{aligned} \boldsymbol{R}_{\mathrm{diag}} &= \boldsymbol{M}^{-1}\boldsymbol{R}\boldsymbol{M} \\ &= \frac{1}{\sqrt{2}}\begin{pmatrix} 1 & 1 \\ -1 & 1 \end{pmatrix}\begin{pmatrix} 1 & r \\ r & 1 \end{pmatrix}\frac{1}{\sqrt{2}}\begin{pmatrix} 1 & -1 \\ 1 & 1 \end{pmatrix} = \begin{pmatrix} 1+r & 0 \\ 0 & 1-r \end{pmatrix} . \end{aligned} \tag{A.7}$$

Note that $\mathrm{Tr}(\boldsymbol{R}_{\mathrm{diag}}) = 2$ and $\det(\boldsymbol{R}_{\mathrm{diag}}) = 1 - r^2$, i.e., the trace and determinant of $\boldsymbol{R}$ remain **invariant** under the diagonalising transformation.

The concepts of eigenvalues (principle components) and eigenvectors (principle axes), as well as of diagonalisation of matrices, generalise in a straightforward though computationally more demanding fashion to arbitrary real-valued **correlation matrices** $\boldsymbol{R} \in \mathbb{R}^{m\times m}$, with $m \in \mathbb{N}$.

[1] Alternatively one can write

$$\boldsymbol{M} = \begin{pmatrix} \cos(\pi/4) & -\sin(\pi/4) \\ \sin(\pi/4) & \cos(\pi/4) \end{pmatrix} ,$$

thus emphasising the character of a rotation of $\boldsymbol{R}$ by an angle $\varphi = \pi/4$.

Appendix B

Distance measures in Statistics

Statistics employs a number of different measures of **distance** d_{ij} to quantify the separation in an m–D space of metrically scaled statistical variables $X, Y, \ldots, Z$ of two statistical units i and j ($i, j = 1, \ldots, n$). Note that, by construction, these measures d_{ij} exhibit the properties $d_{ij} \geq 0$, $d_{ij} = d_{ji}$ and $d_{ii} = 0$. In the following, X_{ik} is the entry of the data matrix $\boldsymbol{X} \in \mathbb{R}^{n \times m}$ relating to the ith statistical unit and the kth statistical variable, etc. The d_{ij} define the elements of a $(\boldsymbol{n} \times \boldsymbol{n})$ **proximity matrix** $\boldsymbol{D} \in \mathbb{R}^{n \times n}$.

Euclidian distance (dimensionful)

This most straightforward, dimensionful distance measure is named after the ancient Greek (?) mathematician Euclid of Alexandria (ca. 325BC–ca. 265BC). It is defined by

$$\boxed{d_{ij}^{E} := \sqrt{\sum_{k=1}^{m} \sum_{l=1}^{m} (X_{ik} - X_{jk}) \delta_{kl} (X_{il} - X_{jl})} \; ,} \tag{B.1}$$

where δ_{kl} denotes the elements of the unit matrix $\mathbf{1} \in \mathbb{R}^{m \times m}$; cf. Ref. [12, Eq. (2.2)].

Mahalanobis distance (dimensionless)

A more sophisticated, **scale-invariant** distance measure in **Statistics** was devised by the Indian applied statistician Prasanta Chandra Mahalanobis (1893–1972); cf. Mahalanobis (1936) [35]. It is defined by

$$\boxed{d_{ij}^{M} := \sqrt{\sum_{k=1}^{m} \sum_{l=1}^{m} (X_{ik} - X_{jk}) S_{kl}^{-1} (X_{il} - X_{jl})} \; ,} \tag{B.2}$$

where S_{kl}^{-1} denotes the elements of the inverse covariance matrix $\boldsymbol{S}^{-1} \in \mathbb{R}^{m \times m}$ relating to $X, Y, \ldots, Z$; cf. Subsec. 4.2.1.

Appendix C

Glossary of technical terms (GB – D)

A
ANOVA: Varianzanalyse
arithmetical mean: arithmetischer Mittelwert
association: Zusammenhang, Assoziation
attribute: Ausprägung, Eigenschaft

B
bar chart: Balkendiagramm
Bayes' theorem: Satz von Bayes
best-fit model: Anpassungsmodell
binomial coefficient: Binomialkoeffizient
bivariate: bivariat, zwei variable Größen betreffend

C
category: Kategorie
causal relationship: Kausalbeziehung
census: statistische Vollerhebung
central limit theorem: Zentraler Grenzwertsatz
certain event: sicheres Ereignis
class interval: Ausprägungsklasse
cluster random sample: Klumpenzufallsstichprobe
coefficient of determination: Bestimmtheitsmaß
coefficient of variation: Variationskoeffizient
combination: Kombination
combinatorics: Kombinatorik
compact: geschlossen, kompakt
concentration: Konzentration
conditional probability: bedingte Wahrscheinlichkeit
confidence interval: Konfidenzintervall
contingency table: Kontingenztafel
continuous data: stetige Daten
convexity: Konvexität

correlation matrix: Korrelationsmatrix
covariance matrix: Kovarianzmatrix
cumulative distribution function (`cdf`): theoretische Verteilungsfunktion

D
data matrix: Datenmatrix
deductive method: deduktive Methode
degrees of freedom: Freiheitsgrade
dependent variable: abhängige Variable
descriptive statistics: Beschreibende Statistik
deviation: Abweichung
direction: Richtung
discrete data: diskrete Daten
disjoint events: disjunkte Ereignisse, einander ausschließend
dispersion: Streuung
distance: Abstand
distortion: Verzerrung
distribution: Verteilung
distributional properties: Verteilungseigenschaften

E
elementary event: Elementarereignis
empirical cumulative distribution function: empirische Verteilungsfunktion
estimator: Schätzer
Euclidian distance: Euklidischer Abstand
event: Ereignis
event space: Ereignisraum
expectation value: Erwartungswert

F
factorial: Fakultät
falsification: Falsifikation
five number summary: Fünfpunktzusammenfassung
frequency: Häufigkeit

G
Gini coefficient: Ginikoeffizient
goodness-of-the-fit: Anpassungsgüte

H
Hessian matrix: Hesse'sche Matrix
histogram: Histogramm
homoscedasticity: Homoskedastizität, homogene Varianz
hypothesis: Hypothese, Behauptung, Vermutung

I
independent variable: unabhängige Variable
inductive method: induktive Methode

inferential statistics: Schließende Statistik
interaction: Wechselwirkung
intercept: Achsenabschnitt
interquartile range: Quartilsabstand
interval scale: Intervallskala
impossible event: unmögliches Ereignis

J
joint distribution: gemeinsame Verteilung

K
$k\sigma$–rule: $k\sigma$–Regel
kurtosis: Wölbung

L
latent variable: latente Variable, Konstrukt
law of large numbers: Gesetz der großen Zahlen
law of total probability: Satz von der totalen Wahrscheinlichkeit
linear regression analysis: lineare Regressionsanalyse
Lorenz curve: Lorenzkurve

M
Mahalanobis distance: Mahalanobis'scher Abstand
manifest variable: manifeste Variable, Observable
marginal frequencies: Randhäufigkeiten
measurement: Messung, Datenaufnahme
median: Median
metrical: metrisch
mode: Modalwert

N
nominal: nominal

O
observable: beobachtbare/messbare Variable, Observable
observation: Beobachtung
operationalisation: Operationalisieren, latente Variable messbar gestalten
opinion poll: Meinungsumfrage
ordinal: ordinal
outlier: Ausreißer

P
p–value: p–Wert
partition: Zerlegung, Aufteilung
pie chart: Kreisdiagramm
point estimator: Punktschätzer
population: Grundgesamtheit
power: Teststärke

power set: Potenzmenge
principle component analysis: Hauptkomponentenanalyse
probability: Wahrscheinlichkeit
probability density function (`pdf`): Wahrscheinlichkeitsdichte
probability function: Wahrscheinlichkeitsfunktion
proximity matrix: Distanzmatrix

Q

quantile: Quantil
quartile: Quartil

R

random sample: Zufallsstichprobe
random experiment: Zufallsexperiment
random variable: Zufallsvariable
range: Spannweite
rank: Rang
ratio scale: Verhältnisskala
raw data set: Datenurliste
realisation: Realisation, konkreter Messwert für eine Zufallsvariable
regression analysis: Regressionsanalyse
regression coefficient: Regressionskoeffizient
rejection region: Ablehnungsbereich
research question: Forschungsfrage
residual: Residuum, Restgröße
risk: Risiko (berechenbar)

S

σ–algebra: σ–Algebra
sample: Stichprobe
sample correlation coefficient: Stichprobenkorrelationskoeffizient
sample covariance: Stichprobenkovarianz
sample mean: Stichprobenmittelwert
sample space: Ergebnismenge
sample variance: Stichprobenvarianz
sampling distribution: Stichprobenverteilung
sampling error: Stichprobenfehler
sampling frame: Auswahlgesamtheit
sampling unit: Stichprobeneinheit
scale-invariant: skaleninvariant
scale level: Skalenniveau
scatter plot: Streudiagramm
shift theorem: Verschiebungssatz
significance: Signifikanz
significance level: Signifikanzniveau
simple random sample: einfache Zufallsstichprobe

skewness: Schiefe
slope: Steigung
spectrum of values: Wertespektrum
spurious correlation: Scheinkorrelation
standard error: Standardfehler
standardisation: Standardisierung
statistical independence: statistische Unabhängigkeit
statistical unit: Erhebungseinheit
statistical variable: Merkmal, Variable
stochastic independence: stochastische Unabhängigkeit
stratified random sample: geschichtete Zufallsstichprobe
strength: Stärke
survey: statistische Erhebung, Umfrage

T
test statistic: Teststatistik, statistische Effektmessgröße
type I error: Fehler 1. Art
type II error: Fehler 2. Art

U
unbiased: erwartungstreu
uncertainty: Unsicherheit (nicht berechenbar)
univariate: univariat, eine variable Größe betreffend
urn model: Urnenmodell

V
value: Wert
variance: Varianz
variation: Variation

W
weighted mean: gewichteter Mittelwert

Z
z–scores: z–Werte

Bibliography

[1] T Bayes (1763) An essay towards solving a problem in the doctrine of chances *Phil. Trans.* **53** 370–418

[2] P L Bernstein (1998) *Against the Gods — The Remarkable Story of Risk* (New York: Wiley) ISBN–10: 0471295639

[3] J–P Bouchaud and M Potters (2003) *Theory of Financial Risk and Derivative Pricing* 2nd Edition (Cambridge: Cambridge University Press) ISBN–13: 9780521741866

[4] J Bortz (2005) *Statistik für Human– und Sozialwissenschaftler* 6th Edition (Berlin: Springer) ISBN–13: 9783540212713

[5] J Bortz and N Döring (2006) *Forschungsmethoden und Evaluation für Human– und Sozialwissenschaftler* 4th Edition (Berlin: Springer) ISBN–13: 9783540333050

[6] K Bosch (1999) *Grundzüge der Statistik* 2nd Edition (München: Oldenbourg) ISBN–10: 3486252593

[7] A Bravais (1846) Analyse mathématique sur les probabilités des erreurs de situation d'un point *Mémoires présentés par divers savants à l'Académie royale des sciences de l'Institut de France* **9** 255–332

[8] H Cramér (1946) *Mathematical Methods of Statistics* (Princeton, NJ: Princeton University Press) ISBN–10: 0691080046

[9] L J Cronbach (1951) Coefficient alpha and the internal structure of tests *Psychometrika* **16** 297–334

[10] P Dalgaard (2008) *Introductory Statistics with R* 2nd Edition (New York: Springer) ISBN–13: 9780387790534

[11] C Duller (2007) *Einführung in die Statistik mit EXCEL und SPSS* 2nd Edition (Heidelberg: Physica) ISBN–13: 9783790819113

[12] H van Elst (2009–2012) *0.1.1 EMQM: Einführung in das Management und seine quantitativen Methoden (Wirtschaftsmathematik)* Vorlesungsskript (Karlsruhe: Karlshochschule International University)

[13] W Feller (1951) The asymptotic distribution of the range of sums of independent random variables *Ann. Math. Statist.* **22** 427–432

[14] R A Fisher (1918) The correlation between relatives on the supposition of Mendelian inheritance *Trans. Roy. Soc. Edinburgh* **52** 399–433

[15] R A Fisher (1924) On a distribution yielding the error functions of several well known statistics *Proc. Int. Cong. Math. Toronto* **2** 805–813

[16] R A Fisher (1935) The logic of inductive inference *J. Roy. Stat. Soc.* **98** 39–82

[17] C F Gauß (1809) Theoria motus corporum coelestium in sectionibus conicis solem ambientum

[18] I Gilboa (2009) *Theory of Decision under Uncertainty* (Cambridge: Cambridge University Press) ISBN–13: 9780521571324

[19] J Gleick (1987) *Chaos — Making a New Science* $n^{\rm th}$ Ed. 1998 (London: Vintage) ISBN–13: 9780749386061

[20] R Hatzinger and H Nagel (2009) *PASW Statistics — Statistische Methoden und Fallbeispiele* (München: Pearson Studium) ISBN–13: 9783827372734

[21] J Hartung, B Elpelt and K–H Klösener (2005) *Statistik: Lehr– und Handbuch der angewandten Statistik* $14^{\rm th}$ Edition (München: Oldenburg) ISBN–10: 3486578901

[22] S Holm (1979) A simple sequentially rejective multiple test procedure *Scand. J. Statist.* **6** 65–70

[23] M Keuls (1952) The use of the "studentized range" in connection with an analysis of variance *Euphytica* **1** 112–122

[24] A Kolmogoroff (1933) *Grundbegriffe der Wahrscheinlichkeitsrechnung* (Berlin: Springer) $2^{\rm nd}$ reprint: (1973) (Berlin: Springer) ISBN–13: 9783540061106

[25] A N Kolmogorov (1933) Sulla determinazione empirica di una legge di distribuzione *Inst. Ital. Atti. Giorn.* **4** 83–91

[26] W H Kruskal and W A Wallis (1952) Use of ranks on one-criterion variance analysis *J. Am. Stat. Assoc.* **47** 583–621

[27] P S Laplace (1812) *Théorie Analytique des Probabilités* (Paris: Courcier)

[28] E L Lehman and G Casella (1998) *Theory of Point Estimation* $2^{\rm nd}$ Edition (New York: Springer) ISBN–13: 9780387985022

[29] H Levene (1960) Robust tests for equality of variances *Contributions to Probability and Statistics: Essays in Honor of Harold Hotelling* eds I Olkin *et al* (Stanford, CA: Stanford University Press) 278–292

[30] J A Levin, J A Fox and D R Forde (2010) *Elementary Statistics in Social Research* 11th Edition (München: Pearson Education) ISBN–13: 9780205636921

[31] R Likert (1932) A technique for the measurement of attitudes *Archives of Psychology* **140** 1–55

[32] J W Lindeberg (1922) Eine neue Herleitung des Exponentialgesetzes in der Wahrscheinlichkeitsrechnung *Math. Z.* **15** 211–225

[33] R Lupton (1993) *Statistics in Theory and Practice* (Princeton, NJ: Princeton University Press) ISBN–13: 9780691074290

[34] A M Lyapunov (1901) Nouvelle forme du théorème sur la limite de la probabilité *Mémoires de l'Académie Impériale des Sciences de St.-Pétersbourg VIIIe Série, Classe Physico–Mathématique* **12** 1–24 [in Russian]

[35] P C Mahalanobis (1936) On the generalized distance in statistics *Proc. Nat. Inst. Sci. India (Calcutta)* **2** 49–55

[36] H B Mann and D R Whitney (1947) On a test of whether one of two random variables is stochastically larger than the other *Ann. Math. Statis.* **18** 50–60

[37] D Newman (1939) The distribution of range in samples from a normal population, expressed in terms of an independent estimate of standard deviation *Biometrika* **31** 20–30

[38] J Neyman and E S Pearson (1933) On the problem of the most efficient tests of statistical hypotheses *Phil. Trans. R. Soc. Lond. A* **231** 289–337

[39] V Pareto (1896) *Cours d'Économie Politique* (Geneva: Droz)

[40] K Pearson (1900) On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling *Phil. Mag. Ser. 5* **50** 157–175

[41] K Pearson (1901) LIII. On lines and planes of closest fit to systems of points in space *Phil. Mag. Ser. 6* **2** 559–572

[42] K Pearson (1903) The law of ancestral heredity *Biometrika* **2** 211–228

[43] R Penrose (2004) *The Road to Reality — A Complete Guide to the Laws of the Universe* 1st Edition (London: Jonathan Cape) ISBN–10: 0224044478

[44] K R Popper (2002) *Conjectures and Refutations: The Growth of Scientific Knowledge* 2nd Edition (London: Routledge) ISBN–13: 9780415285940

[45] H Rinne (2008) *Taschenbuch der Statistik* 4th Edition (Frankfurt/Main: Harri Deutsch) ISBN–13: 9783817118274

[46] H Scheffé (1959) *The Analysis of Variance* (New York: Wiley)
Reprint: (1999) (New York: Wiley) ISBN-13: 9780471345053

[47] D S Sivia and J Skilling (2006) *Data Analysis — A Bayesian Tutorial* 2nd Edition (Oxford: Oxford University Press) ISBN–13: 9780198568322

[48] N Smirnov (1939) On the estimation of the discrepancy between empirical curves of distribution for two independent samples *Bull. Math.Univ. Moscou* **2** fasc. 2

[49] L Smith (2007) *Chaos — A Very Short Introduction* (Oxford: Oxford University Press) ISBN–13: 9780192853783

[50] G W Snedecor (1934) *Calculation and Interpretation of Analysis of Variance and Covariance* (Ames, IA: Collegiate Press)

[51] C Spearman (1904) The proof and measurement of association between two things *Am. J. Psych.* **15** 72–101

[52] Student [W S Gosset] (1908) The probable error of a mean *Biometrika* **6** 1–25

[53] `sueddeutsche.de` (2012) Reiche trotz Finanzkrise immer reicher URL (cited on September 19, 2012): www.sueddeutsche.de/wirtschaft/neuer-armuts-und-reichtumsbericht-der-bundesregierung-reiche-trotz-f

[54] E Svetlova and H van Elst (2012) How is non-knowledge represented in economic theory? *Preprint* arXiv:1209.2204v1 [q-fin.GN]

[55] N N Taleb (2007) *The Black Swan — The Impact of the Highly Improbable* (London: Penguin) ISBN–13: 9780141034591

[56] H Toutenburg (2004) *Deskriptive Statistik* 4th Edition (Berlin: Springer) ISBN–10: 3540222332

[57] H Toutenburg (2005) *Induktive Statistik* 3rd Edition (Berlin: Springer) ISBN–10: 3540242937

[58] W M K Trochim (2006) *Web Center for Social Research Methods* URL (cited on June 22, 2012): `www.socialresearchmethods.net`

[59] J W Tukey (1977) *Exploratory Data Analysis* (Reading, MA: Addison–Wesley) ISBN–10: 0201076160

[60] M C Wewel (2008) *Statistik im Bachelor–Studium der BWL und VWL* 2nd Edition (München: Pearson Studium) ISBN–13: 9783827372246

[61] K Wiesenfeld (2001) Resource Letter: ScL-1: Scaling laws *Am. J. Phys.* **69** 938–942

[62] F Wilcoxon (1945) Individual comparisons by ranking methods *Biometrics Bulletin* **1** 80–83